

Explanatory and Predictive Model for Analyzing University Entrance Scores Using Linear Regression

Edwar Saire-Peralta¹; Sonia Calloapaza-Pari¹; Ricardo Calienes-Rodríguez¹; Rene Nieto-Valencia¹; Christian Revilla-Arroyo¹

¹Universidad Nacional de San Agustín de Arequipa, Perú

esaire@unsa.edu.pe, scallopapaza@unsa.edu.pe, rcalienesr@unsa.edu.pe, rnietov@unsa.edu.pe, crevilla@unsa.edu.pe

Abstract - This research aims to analyze university entrance scores using linear regression models from both explanatory and predictive perspectives. A quantitative cross-sectional design was used with 545 university students, considering socioeconomic, academic, motivational, and demographic variables. An exploratory data analysis was conducted, revealing weak but statistically significant associations between entrance scores and variables such as income and years of study. Subsequently, an explanatory model was created using classical linear regression (OLS), incorporating a process for selecting the most significant variables. The final model yielded an adjusted R^2 of 0.126, indicating limited explanatory power. The variables that showed statistically significant effects were years of study, type of high school attended, and the mother's educational level. Next, a predictive approach based on linear regression was implemented under the machine learning paradigm, using validation with training and test data, as well as automatic variable selection via LASSO regression. The performance metrics showed low predictive capacity, reflecting limitations in the model's generalizability. The final results of the study show that linear regression is useful for explaining significant associations but has predictive limitations; therefore, both approaches complement each other to offer a comprehensive perspective on university admissions.

Keywords-- Linear regression, Explanatory model, Predictive model, University entrance score, Factors associated with performance

Modelo explicativo y predictivo para el análisis del puntaje de ingreso a la universidad mediante regresión lineal

Edwar Saire-Peralta¹ ; Sonia Calloapaza-Pari¹ ; Ricardo Calienes-Rodríguez¹ ; Rene Nieto-Valencia¹ ; Christian Revilla-Arroyo¹ 

¹Universidad Nacional de San Agustín de Arequipa, Perú

esaire@unsa.edu.pe, scalloapaza@unsa.edu.pe, rcalienesr@unsa.edu.pe, rnietov@unsa.edu.pe, crevilla@unsa.edu.pe

Resumen- La presente investigación tiene como objetivo analizar el puntaje de ingreso a la universidad mediante el uso de modelos de regresión lineal desde el enfoque explicativo y desde el enfoque predictivo. Se utilizó un diseño transversal cuantitativo, con 545 estudiantes universitarios, considerando variables socioeconómicas, académicas, motivacionales y demográficas. Se inició con un análisis exploratorio de los datos, donde los resultados mostraron asociaciones débiles, pero estadísticamente significativas entre el puntaje de ingreso y variables como el ingreso y el tiempo de estudio. Posteriormente, creo un modelo explicativo mediante regresión lineal clásica (OLS), incorporando un proceso de selección de variables más significativas, donde el modelo final obtuvo un R^2 ajustado de 0.126, lo que indica una capacidad explicativa limitada, donde las variables que mostraron efectos estadísticamente significativos se encontraron el tiempo de estudio, el tipo de colegio de procedencia y el nivel educativo de la madre. Luego, se implementó un enfoque predictivo basado en regresión lineal bajo el paradigma de aprendizaje automático, utilizando validación con datos de entrenamiento y prueba, así como selección automática de variables mediante regresión LASSO, donde los resultados de las métricas de desempeño mostraron una baja capacidad predictiva, lo que refleja limitaciones en la generalización del modelo. Los resultados finales del estudio muestran que la regresión lineal es útil para explicar asociaciones significativas, pero tiene limitaciones predictivas, por lo que ambos enfoques se complementan para ofrecer una perspectiva integral de ingreso a la universidad.

Palabras clave: Regresión lineal, Modelo explicativo, Modelo predictivo, Puntaje de ingreso universitario, Factores asociados al rendimiento.

I. INTRODUCCIÓN

El rendimiento académico es un concepto multidimensional que se considera un indicador sólido del éxito educativo y la permanencia estudiantil en la educación superior [1]. Muchas investigaciones han mostrado que el rendimiento académico está influenciado por factores socioeconómicos, académicos, motivacionales y demográficos, como el nivel educativo de los padres, ingreso familiar, hábitos de estudio, motivación y género del estudiante [1], [2], [3]. Dichos factores no actúan de manera aislada, sino que

interactúan entre sí, reflejando la complejidad del proceso educativo y la necesidad de abordarlo desde múltiples dimensiones [4], [5].

Desde un enfoque de investigación, la regresión lineal se ha utilizado ampliamente para explicar cómo las variables independientes se relacionan con resultados de valores continuos, como calificaciones o puntajes de ingreso universitario. Dicho enfoque permite estimar coeficientes, evaluar la significancia estadística de las variables predictores y construir un modelo explicativo sólido, útil para comprender los determinantes del rendimiento académico [6]. Asimismo, los métodos basados en regresión lineal para predecir, incluidos las técnicas de selección de variables regularizada como LASSO (Least Absolute Shrinkage and Selection Operator), se utilizan para anticipar el rendimiento de nuevos estudiantes en base a patrones aprendidos de los datos, evaluando su capacidad de generalización mediante métricas como RMSE, MAE y R^2 [7], [8].

La combinación de los enfoques, explicativo y predictivo, permite un análisis más completo del puntaje de ingreso universitario. Por un lado, el modelo explicativo permite identificar relaciones significativas y aportar resultados sobre los determinantes estadísticos del rendimiento, y por otro lado se tiene el modelo predictivo que permite evaluar la capacidad de anticipación y generalización de esas relaciones sobre nuevos casos. Esta complementariedad brinda una visión más robusta y sólida de los factores que influyen en el rendimiento académico [9], [10].

En este contexto, el estado del arte señala que muchos estudios han abordado el rendimiento académico usando regresión lineal o técnicas predictivas de manera aislada [10], [11]. Muchas investigaciones en América Latina y otras regiones han documentado asociaciones significativas entre variables socioeconómicas, familiares y académicas con el desempeño universitario [10], [11], pero por otro lado existe una brecha clara, el cual es que existen pocos trabajos que han comparado la capacidad explicativa de un modelo de regresión lineal clásica con la capacidad predictiva de un modelo basado en aprendizaje automático sobre el mismo conjunto de datos y

con las mismas variables [12], [13]. También se ha encontrado que la mayoría de los estudios analiza dimensiones aisladas, sin considerar cómo interactúan de manera conjunta factores socioeconómicos, académicos, motivacionales y demográficos [14], [15], lo que limita la comprensión integral del rendimiento académico.

En este contexto particular, el presente estudio tiene como objetivo analizar y complementar un modelo explicativo y un modelo predictivo basados en regresión lineal para estimar el puntaje de ingreso a la universidad, donde se consideran variables socioeconómicas, académicas, motivacionales y demográficas. Con ambos modelos, este trabajo busca aportar evidencia sólida que identifique las variables más significativas del rendimiento académico y, al mismo tiempo, evalúe la capacidad de predicción de estos determinantes en nuevos casos, contribuyendo a cerrar la brecha metodológica existente y ofreciendo información útil para la toma de decisiones educativas.

II. METODOLOGÍA

A. Diseño y enfoque de la investigación

La presente investigación se desarrolló bajo un enfoque cuantitativo, debido a la medición, análisis y modelado de relaciones estadísticas entre variables sociodemográficas, académicas, familiares y motivacionales y el puntaje de ingreso a la universidad. El diseño de la investigación es no experimental de corte transversal, ya que los datos fueron recolectados en un solo momento temporal y no se manipularon deliberadamente las variables independientes. El estudio se ubica en un nivel correlacional-explicativo y predictivo, pues busca tanto identificar asociaciones significativas entre las variables como evaluar la capacidad de los modelos de regresión lineal para explicar y predecir el puntaje de ingreso, utilizando enfoques tradicionales y de aprendizaje automático.

B. Participantes

La población estuvo conformada por estudiantes que postularon a una universidad pública peruana. La muestra utilizada incluyó 545 estudiantes, seleccionados mediante un muestreo no probabilístico por conveniencia, debido a la disponibilidad de datos completos en las variables de interés. Los datos que se recopilaban por cada estudiante se agruparon en características sociodemográficas (sexo, edad, zona de residencia), académicas (tipo de colegio, modalidad de ingreso, preparación previa), familiares y socioeconómicas (nivel educativo de los padres, ingreso familiar, clima familiar) y motivacionales (motivación, hábitos y tiempo de estudio).

C. Instrumento

Los datos se recolectaron mediante un cuestionario estructurado, complementado con información oficial del puntaje de ingreso a la universidad. El cuestionario fue diseñado para capturar información detallada de las cuatro

dimensiones consideradas (sociodemográfica, académica, familiar-socioeconómica y motivacional). Todas las variables fueron codificadas numéricamente para facilitar el análisis estadístico y garantizar compatibilidad con los algoritmos de regresión lineal. La validez del instrumento se basó en estudios previos y la confiabilidad se verificó mediante análisis preliminares de consistencia interna.

D. Procedimiento

El procedimiento de recolección de datos incluyó la solicitud del puntaje de ingreso a la universidad y la aplicación del cuestionario a los postulantes. En total se recolectaron datos en base a 22 variables, donde la variable dependiente es puntaje y las variables independientes se muestra en la Tabla I distribuidas en dimensiones socioeconómicas, académicas, motivacionales y demográficas.

TABLA I
VARIABLES INDEPENDIENTES

Variable	Tipo	Dimensión
sexo	Categoría	Demográfica
edad	Categoría	Demográfica
auto_motivacion	Ordinal	Motivacional
tiempo_estudio	Ordinal	Motivacional
vida_academica	Ordinal	Académica
procrastinacion	Ordinal	Académica
asistencia	Ordinal	Académica
deportes	Categoría	Académica
celular_ocio	Ordinal	Motivacional
ocupacion_madre	Categoría	Socioeconómica
ocupacion_padre	Categoría	Socioeconómica
financia_estudios	Categoría	Socioeconómica
ingreso_familiar	Categoría	Socioeconómica
violencia_familiar	Categoría	Socioeconómica
interes_padres	Ordinal	Motivacional
eleccion_carrera	Categoría	Académica
trabaja_estudio	Categoría	Académica
tipo_colegio	Categoría	Académica
modalidad_ingreso	Categoría	Académica
instruccion_padre	Categoría	Socioeconómica
instruccion_madre	Categoría	Socioeconómica

. Los datos recolectados fueron preprocesados antes del análisis, incluyendo la limpieza de valores faltantes, codificación de variables categóricas y normalización de variables continuas según correspondiera. Se eliminaron variables con un porcentaje elevado de datos faltantes que pudieran afectar la robustez del análisis, como fue el caso de “violencia familiar”. Seguidamente todos los datos fueron utilizados para generar el modelo explicativo, sin embargo, para generar el modelo predictivo los datos fueron divididos en conjuntos de entrenamiento y prueba.

E. Análisis estadístico de los datos

El análisis de los datos se desarrolló en dos fases complementarias. La primera fase corresponde al modelo explicativo, el cual se implementó una regresión lineal múltiple clásica (Ordinary Least Squares, OLS) para evaluar la relación entre las variables independientes y el puntaje de

ingreso. Antes de generar el modelo se verificaron los supuestos de normalidad, homocedasticidad, linealidad e independencia de los errores. Además, se evaluó la multicolinealidad mediante el VIF y se aplicó un proceso de selección de variables paso a paso para identificar aquellas con efectos significativos ($p < 0.05$). Se reportaron los coeficientes estandarizados, R^2 ajustado y valores p para interpretar el aporte de cada variable.

La segunda fase fue implementar el modelo predictivo, donde también se empleó el algoritmo de regresión lineal desde un enfoque de aprendizaje automático, utilizando partición de datos en entrenamiento (80%) y prueba (20%) y escalado de variables mediante StandardScaler. Para la selección de variables más relevantes se aplicó LASSO, que permite reducir coeficientes de variables menos importantes a cero y mejorar la generalización del modelo. El desempeño predictivo se evaluó mediante métricas de error como RMSE, MAE y R^2 . Este enfoque complementa al modelo explicativo, ya que permite anticipar el rendimiento de nuevos estudiantes basándose en patrones aprendidos sin limitarse a inferencias estrictamente estadísticas.

La creación y análisis de los modelos fue realizado utilizando Python, con bibliotecas como Pandas, NumPy, Statsmodels y Scikit-learn, que facilitaron tanto la implementación de los modelos como la obtención de métricas de desempeño y visualización de resultados.

III. RESULTADOS

A. Resultados descriptivos

Al hacer un análisis descriptivo se mostró que el puntaje de ingreso a la universidad presentó una media de 54.19 con una desviación estándar de 9.65, lo que indica una variabilidad moderada entre los estudiantes. Los valores del puntaje oscilaron entre 20 y 81 puntos, con una mediana de 53, donde el 50% de los estudiantes obtuvo puntajes entre 48 (percentil 25) y 60 puntos (percentil 75). El diagrama de caja (boxplot) como se muestra en la fig. 1, complementa esta descripción, visualizando claramente la mediana, la dispersión intercuartílica y los posibles valores atípicos. Se observan algunos estudiantes con puntajes extremos, algunos muy bajos y otros muy altos, lo que sugiere heterogeneidad en el rendimiento académico de la muestra, además se muestra la presencia de casos fuera del rango central que podrían influir en los análisis posteriores.

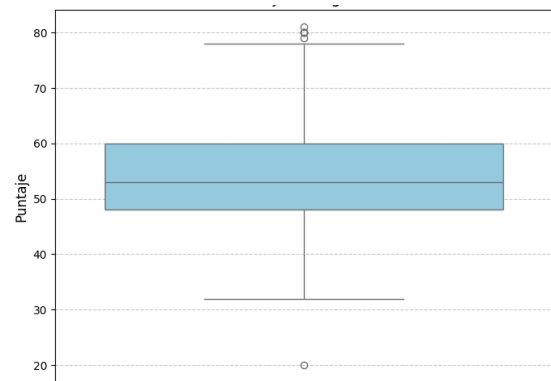


Fig. 1 Distribución del puntaje de ingreso a la universidad

Con el fin de resumir la información y facilitar la interpretación de los resultados, en la Tabla II se presentan algunas de las variables categóricas del estudio. Para cada variable se indica la categoría que presenta la mayor frecuencia dentro de la muestra. Las demás variables se incluyen en el análisis de los modelos explicativo y predictivo, pero no se muestran aquí para evitar saturar la sección de resultados.

TABLA II
VARIABLES CATEGORICAS DE ESTUDIO

Variable	Categoría principal	Porcentaje (%)
Sexo	Femenino	60.18
Edad	18–19 años	41.83
Auto-motivación	Sí	87.89
Tiempo de estudio	2–3 horas	35.96
Vida académica	Bueno	50.46
Tipo de colegio	Público	60.00
Instrucción madre	Secundaria	36.51
Instrucción padre	Secundaria	40.18
Modalidad de ingreso	Examen ordinario	61.65

B. Análisis exploratorio de relaciones bivariadas

Se ha desarrollado un análisis preliminar para identificar relaciones entre el puntaje de ingreso y las variables independientes. En la Tabla III se muestra las correlaciones significativas (Pearson) para variables ordinales y binarias.

TABLA III
CORRELACION DE VARIABLES

Variable	r	p-valor
ingreso_familiar	0.149	<0.001
tiempo_estudio	0.099	0.020

Otras variables ordinales como procrastinación, asistencia, sexo, deportes, interés de los padres, edad, automotivación, vida académica, celular ocio y violencia familiar no mostraron correlaciones significativas ($p > 0.05$).

Para evaluar la relación entre las variables categóricas nominales (dummy) y el puntaje de ingreso a la universidad, se aplicó un ANOVA univariado como se muestra en la Tabla

IV. El análisis ha permitido determinar si existen diferencias significativas en la media del puntaje entre los distintos grupos de cada variable categórica. Cada variable dummy representa la presencia o ausencia de una categoría específica, por lo que el ANOVA permite identificar qué categorías muestran un efecto estadísticamente significativo sobre la variable dependiente. Las variables que presentan un p-valor menor a 0.05 se consideran candidatas fuertes para el modelo explicativo, ya que sugieren que la pertenencia a esa categoría influye de manera relevante en el puntaje universitario.

TABLA IV
RESULTADOS DE ANOVA

Variable	p-valor
tipo_colegio_Público	1.92e-13
tipo_colegio_Privado	3.06e-10
instruccion_padre_Secundaria	9.46e-05
instruccion_madre_Pregado	1.33e-04
instruccion_madre Primaria	5.75e-04
instruccion_padre_Pregado	5.98e-04
ocupacion_madre_Profesional	6.56e-04

C. Resultados del modelo explicativo (OLS)

Se implementó un modelo de regresión lineal múltiple con backward elimination para obtener solo variables significativas. El modelo final incluyó cuatro variables tal como se muestra en la Tabla V.

TABLA V
VARIABLES MAS SIGNIFICATIVAS

Variable	Coef.	Err. Est.	p-valor
const	57.06	1.10	<0.001
tiempo_estudio	0.80	0.33	0.016
tipo_colegio_Público	-5.09	0.84	<0.001
instruccion_madre Primaria	-3.25	1.12	0.004
instruccion_madre_Secundaria	-3.39	0.88	<0.001

El valor de R^2 igual a 0.132 y R^2 ajustado igual a 0.126, señalan que el modelo explica aproximadamente el 12.6% de la variabilidad del puntaje de ingreso. Todas las variables finales son estadísticamente significativas ($p < 0.05$) y no presentan problemas de multicolinealidad ($VIF < 5$ para todas). Con respecto a la variable tiempo de estudio, se encontró que mientras mayor dedicación al estudio se asocia con puntajes más altos. El tipo de colegio público se relaciona con un puntaje promedio más bajo respecto a otros tipos de colegios y el nivel educativo de la madre (Primaria y Secundaria). Con este conjunto de datos se señala que tener madres con educación primaria o secundaria se asocia con una disminución en el puntaje de ingreso, indicando influencia familiar sobre el rendimiento académico inicial.

D. Resultados del modelo predictivo

También se implementó un modelo de regresión lineal, previamente se aplicó LASSO para identificar las variables más relevantes en la predicción del puntaje de ingreso. La aplicación de LASSO ha permitido filtrar y reducir la dimensionalidad del conjunto de datos, descartando variables

con coeficientes cercanos a cero y reteniendo aquellas con mayor influencia sobre el puntaje.

Después de la aplicación de LASSO, se seleccionaron 36 de las 42 variables iniciales. Entre las variables más relevantes se encuentran tipo de colegio público (-1.73), instrucción de la madre secundaria (-1.53), instrucción de la madre primaria (-1.27), tiempo de estudio (1.16), elección de carrera por posibilidad económica (-0.93), procrastinación (0.89), tipo de colegio privado (0.83), sexo (-0.82), instrucción del padre pregrado (0.75), deportes (-0.73), edad (0.71), elección de carrera por demanda laboral (-0.70), modalidad de ingreso examen extraordinario (0.69), asistencia (0.68), vida académica (-0.67), ocupación madre otro (0.54), trabaja estudio no trabajo (0.43), instrucción padre técnico (0.42), financia estudios yo mismo (-0.39) y financia estudios otros parientes (-0.35), entre otras variables con coeficientes menores.

Una vez identificado las variables más significativas a través de LASSO, se implementó el modelo con regresión lineal y el desempeño en el conjunto de prueba mostró RMSE igual a 9.20, MAE igual a 7.05 y R^2 igual a -0.042, lo que indica una capacidad limitada de generalización. Los resultados de LASSO representan las variables más influyentes para la predicción del puntaje de ingreso, destacando principalmente la procedencia escolar, la educación de los padres y las variables académicas y motivacionales del estudiante.

E. Comparación entre modelo explicativo y predictivo.

En la implementación de ambos modelos se utilizó regresión lineal, pero con objetivos distintos, donde el modelo explicativo (OLS) permite evaluar la significancia y el efecto individual de las variables sobre el puntaje de ingreso, mientras que el modelo predictivo se centra en identificar automáticamente las variables más relevantes para anticipar el puntaje de nuevos estudiantes. Cada modelo captura las variables significativas de manera diferente, por un lado, tenemos a OLS donde usa pruebas de significancia estadística, y por otro lado tenemos a LASSO donde usa penalización de coeficientes cercanos a cero.

Los resultados más relevantes muestran que, en OLS, tiempo de estudio (β igual a 0.80, p igual a 0.016), tipo de colegio público (β igual a -5.09, $p < 0.001$) y nivel educativo de la madre primaria y secundaria (β aproximado a -3.3, $p < 0.01$) son las variables que más explican la variabilidad del puntaje, con un R^2 ajustado igual a 0.126. Por su parte, LASSO seleccionó 36 variables de 42 iniciales, destacando principalmente tipo de colegio, instrucción de la madre y tiempo de estudio, con un desempeño predictivo limitado (RMSE igual a 9.20, MAE igual a 7.05, R^2 igual a -0.042).

Para resumir, podemos ver que mientras que OLS proporciona claridad sobre los efectos individuales y su significancia, LASSO permite capturar patrones combinados y anticipar resultados futuros, identificando variables clave que

pueden pasar desapercibidas en un análisis puramente explicativo. La comparación de ambos modelos confirma que ambos enfoques de regresión lineal se complementan: OLS explica relaciones directas, y LASSO identifica los factores más relevantes para la predicción, ofreciendo una visión más integral de los determinantes del puntaje de ingreso.

IV. DISCUSIÓN

Los resultados de esta investigación aportan resultados claros sobre los aportes y limitaciones de los enfoques explicativo y predictivo basados en regresión lineal para analizar el puntaje de ingreso universitario. El uso combinado de regresión lineal tradicional (OLS) con regresión lineal predictiva regularizada sobre el mismo conjunto de datos y las mismas variables constituye una fortaleza metodológica importante, ya que ambos enfoques se complementan, donde OLS permite identificar y cuantificar los efectos individuales de las variables, mientras que el modelo predictivo resalta los patrones combinados de variables para anticipar resultados futuros.

El modelo de regresión lineal (OLS) identificó que el tiempo de estudio, el tipo de colegio de procedencia y el nivel educativo de la madre ejercen efectos significativos sobre el puntaje de ingreso a la universidad. Estos resultados son consistentes con la literatura previa, que destaca cómo tanto el capital humano acumulado mediante el esfuerzo académico [15] como las condiciones socioeconómicas de origen [2] son determinantes clave del éxito en las evaluaciones de alto impacto. La investigación de Sirin (2005) demuestra que el estatus socioeconómico y la educación de los padres están asociados positivamente con el logro académico, reportando tamaños del efecto moderados. Estos resultados son consistentes con la capacidad explicativa de nuestro modelo (R^2 ajustado aproximado a 0.126), sugiriendo que, si bien estas variables son significativas, existen otros factores no observados que también influyen en el puntaje [2].

El uso de LASSO como método de selección automática de variables permitió identificar 36 variables relevantes, integrando factores académicos, motivacionales y demográficos. Este enfoque es consistente con investigaciones recientes [16], [17], las cuales sugieren que las técnicas de regularización son más eficaces que los modelos OLS tradicionales para gestionar conjuntos extensos de predictores, permitiendo descubrir patrones de influencia complejos y mitigar el sobreajuste sin depender únicamente de pruebas de significancia individuales.

Aunque el desempeño predictivo de LASSO mostró limitaciones importantes ($\$RMSE = 9.20$, $MAE = 7.05$, R^2 negativo), este resultado es consistente con la literatura especializada. Revisiones sistemáticas y estudios de minería de datos educativos [18], [19] subrayan que los modelos lineales a menudo son insuficientes para capturar las relaciones no lineales inherentes a los datos académicos y de comportamiento, sugiriendo que algoritmos de mayor

complejidad suelen ofrecer una mayor capacidad predictiva en estos contextos.

Comparado con otros enfoques clásicos, la regresión lineal con LASSO ofrece ventajas críticas de interpretabilidad y selección de variables [20]. Como señalan los referentes del área [21], este método permite anticipar resultados manteniendo la transparencia en los coeficientes, evitando el problema de 'caja negra' común en algoritmos más complejos, lo que facilita la toma de decisiones basada en evidencia.

Desde una perspectiva pedagógica y estadística, la complementariedad metodológica de este estudio —al integrar modelos explicativos y predictivos— permite identificar variables significativas en promedio y, simultáneamente, explorar patrones complejos para predecir puntajes futuros. Esta dualidad metodológica aborda la distinción clásica entre explicar y predecir [9], superando enfoques tradicionales que se limitan a uno de los dos análisis [10] y proporcionando una herramienta más robusta para la gestión educativa.

En suma, los hallazgos respaldan que un enfoque combinado proporciona una comprensión más completa: OLS explica relaciones directas y significativas, mientras que LASSO identifica factores clave para anticipar resultados. Esta dualidad evidencia la utilidad de ambos enfoques como herramientas complementarias en la investigación educativa, permitiendo diseñar intervenciones pedagógicas basadas tanto en el entendimiento de las causas como en la precisión de la detección temprana

V. CONCLUSIONES

El presente estudio demuestra que la aplicación combinada de regresión lineal con enfoque explicativo (OLS) y predictivo (LASSO) sobre el mismo conjunto de datos permite obtener una comprensión integral del puntaje de ingreso universitario. Los resultados muestran que variables como el tiempo de estudio, el tipo de colegio de procedencia y el nivel educativo de la madre son los factores con mayor efecto significativo sobre el puntaje, confirmando la influencia conjunta de factores académicos y familiares en el desempeño inicial de los estudiantes.

La implementación de LASSO permitió seleccionar 36 variables relevantes, incluyendo factores académicos, familiares, motivacionales específicos y demográficos. Estos resultados evidencian que el éxito en el ingreso a la universidad no solo depende de las condiciones socioeconómicas preexistentes, sino también de la actitud y disposición del estudiante detectadas por el modelo predictivo, lo cual identifica patrones combinados que podrían no detectarse mediante un análisis explicativo clásico. Esta selección automática aporta información complementaria al modelo OLS, ofreciendo una perspectiva más completa sobre los determinantes del puntaje.

La unión de ambos enfoques constituye la principal fortaleza metodológica del estudio tratado. Mientras que el OLS cuantifica efectos individuales y establece significancia

estadística, LASSO actúa como un filtro de relevancia que anticipa resultados y resalta patrones combinados. Esta visión dual proporciona una herramienta más robusta para la gestión académica que los análisis tradicionales.

Sin embargo, el modelo predictivo construido con las variables seleccionadas por LASSO es limitado, reflejando la complejidad multifactorial del puntaje de ingreso y la variabilidad individual de los estudiantes. Estos resultados sugieren que las investigaciones futuras podrían explorar modelos más sofisticados o incluir variables dinámicas y longitudinales para mejorar la capacidad de generalización.

Finalmente, este estudio contribuye metodológicamente al mostrar que un enfoque que combina regresión lineal explicativa y predictiva sobre los mismos datos ofrece evidencia complementaria y replicable, superando limitaciones de investigaciones previas que se centraban únicamente en un tipo de análisis o consideraban un número limitado de variables. En conjunto, estos hallazgos proporcionan una herramienta útil para la comprensión de los determinantes del rendimiento académico y para la planificación de estrategias de apoyo educativo.

AGRADECIMIENTO/RECONOCIMIENTO

Se agradece especialmente a los alumnos de la Universidad Nacional de San Agustín, especialmente de las carreras de ingeniería, así como a los colegas que contribuyeron al desarrollo de este estudio.

REFERENCIAS

- [1] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.
- [2] Selcuk R Sirin, "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research," *Rev. Educ. Res.*, vol. 75, no. 3, pp. 417–453, Sep. 2005, doi: 10.3102/00346543075003417.
- [3] Angela L Duckworth and Martin E P Seligman, "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents," *Psychol. Sci.*, vol. 16, no. 12, pp. 939–944, Dec. 2005, doi: 10.1111/j.1467-9280.2005.01641.x.
- [4] A. Wigfield and J. S. Eccles, "Expectancy–Value Theory of Achievement Motivation," *Contemp. Educ. Psychol.*, vol. 25, no. 1, pp. 68–81, 2000, doi: <https://doi.org/10.1006/ceps.1999.1015>.
- [5] Organisation for Economic Co-operation and Development - OECD, *Where all students can succeed.*, vol. II. 2018.
- [6] M. F. Boné Andrade, "Impacto de factores socioeconómicos en el rendimiento académico: análisis con modelos de regresión," *Rev. Multidiscip. Estud. Gen.*, vol. 2, no. 4, pp. 1–9, 2023, doi: 10.70577/reg.v2i4.45.
- [7] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016, pp. 18–20, doi: 10.1109/ICACA.2016.7887916.
- [8] S. B. Narváez, "Análisis Predictivo del Rendimiento Académico en Estudiantes Canarios de Primaria usando Modelos Lasso," UNIVERSIDAD DE LA LAGUNA, 2021.
- [9] G. Shmueli, "To explain or to predict?," *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, 2010, doi: 10.1214/10-STS330.
- [10] G. M. Garbanzo Vargas, "Factores asociados al rendimiento académico tomando en cuenta el nivel socioeconómico: Estudio de regresión múltiple en estudiantes universitarios," *Rev. Electrónica Educ.*, vol. 18, no. 1, pp. 119–154, 2014, doi: 10.15359/ree.18-1.6.
- [11] P. Acosta and P. Pizarro, "Predicción del Rendimiento Académico en la Educación Superior usando Minería de Datos y su Comparación con Técnicas Estadísticas," 2011.
- [12] M. V. García Jiménez, J. M. A. Izquierdo, and A. Jiménez Blanco, "La predicción del rendimiento académico: Regresión lineal versus regresión logística," *Psicothema*, vol. 12, no. SUPPL. 2, pp. 248–252, 2000.
- [13] R. T. Mora García, "Factores que intervienen en el rendimiento académico universitario: Un estudio de caso," *Opcion*, vol. 31, no. Special Issue 6, pp. 1041–1063, 2015.
- [14] C. Augusto and V. Albuja, "Factores socioeconómicos y su relación con el desempeño académico: Una revisión bibliográfica Socioeconomic factors and their relationship with academic performance: A bibliographic review," *Sci. Res. J. Cent. Investig. y Desarro. Intelect. CIDI E*, vol. 4, pp. 43–67, 2024.
- [15] M. M. Dubuc, M. Aubertin-Leheudre, and A. D. Karelis, "Predictors of Academic Performance in High School Students: The Longitudinal ASAP Study," *Int. J. Exerc. Sci.*, vol. 15, no. 4, pp. 616–631, 2022, doi: 10.70252/ZAZQ9758.
- [16] A. G. R. Sandeepa and S. Mohottala, "Evaluation of Machine Learning Models in Student Academic Performance Prediction," *2025 5th Int. Conf. Adv. Res. Comput. Converging Horizons Uniting Discip. Comput. Res. through AI Innov. ICARC 2025 - Proc.*, 2025, doi: 10.1109/ICARC64760.2025.10963104.
- [17] F. Salas and J. Caldas, "Prediciendo el rendimiento académico de estudiantes de pregrado en una universidad destacada de Perú: Una aproximación con herramientas de Machine Learning," *Educación*, vol. 33, no. 64 SE-Monográfico, pp. 55–85, Apr. 2024, doi: 10.18800/educacion.202401.M003.
- [18] Rahul and R. Katarya, "A Systematic Review on Predicting the Performance of Students in Higher Education in Offline Mode Using Machine Learning Techniques," *Wirel. Pers. Commun.*, vol. 133, no. 3, pp. 1643–1674, 2023, doi: 10.1007/s11277-023-10838-x.
- [19] M. Yağci, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
- [20] D. Ying and J. Ma, "Student Performance Prediction with Regression Approach and Data Generation," *Appl. Sci.*, vol. 14, no. 3, 2024, doi: 10.3390/app14031148.
- [21] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning, Springer Texts*, vol. 102. 2006.