MATEMÁTICA
ESPOL - FCNM JOURNAL
2021, Vol. 19, No. 2 Julio - Diciembre

Pronóstico del rendimiento académico en ciencias exactas para admisiones de una universidad pública, utilizando regresión logística binaria

Forecasting of academic performance in exact sciences for a public university admissions using binary logistic regression

Pedro Ramos De Santis

Recepción: 07/03/2021 **Aceptación:** 03/06/2021 **Publicación:** 20/07/2021

Abstract Because of the deficient secondary level education that the applicant receives before entering the university and the academic rigor of the entrance exam as well as the admission course in the Escuela Superior Politécnica del Litoral (ESPOL), there is an adverse scenario for many applicants who delay the admission to the institution or are unable to do so. Identifying and taking advantage of the benefits of using an innovate active learning methodology as an alternative to traditional learning methodology can help to improve this situation. The objective of this article is to predict the academic performance of the ESPOL applicants in exact sciences, emphasizing the differences between those who attend the course with the active learning modality and those who do it with the traditional modality. The data analysis considers the 558 applicants registered in the intensive course in February 2020. A binary logistic regression technique is applied, with a dichotomous dependent variable called academic performance and dependent variables of academic and demographic nature. A relevant of this research indicates that one of the main predictors of academic performance is the modality with which course is attended. Applying this innovative and technological methodology allows a process with a higher admission rate and academic performance compared to the traditional methodology.
Keywords academic performance, binary logistic analysis, learning process.

Resumen Como consecuencia de la deficiente educación secundaria que recibe el postulante antes de ingresar a la Universidad y del rigor académico tanto del examen

Pedro Senatore Ramos De Santis, M.Sc.
Docente, ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ciencias Naturales y Matemáticas (FCNM), Campus Gustavo Galindo Km. 30.5 Vía Perimetral P.O. Box 09-01-5863, Guayaquil, Ecuador, e-mail: pramos@espol.edu.ec,
https://orcid.org/0000-0002-5968-481X
Doctorante, Universidad Nacional de Tumbes, UNTUMBES, Escuela de Post-Grado, Tumbes, Perú

de ingreso como del curso de admisión en la Escuela Superior Politécnica del Litoral (ESPOL), existe un escenario adverso para una gran cantidad de postulantes que retrasan su ingreso a la institución o no logran hacerlo. Identificar y aprovechar los beneficios del uso de una metodología de aprendizaje activo innovadora como alternativa a la metodología de aprendizaje tradicional puede ayudar a mejorar esta situación. El objetivo de este artículo es pronosticar el rendimiento académico del postulante a la ESPOL en el área de ciencias exactas, enfatizando las diferencias entre quienes atienden el curso con la modalidad de aprendizaje activo y los que lo hacen con la modalidad tradicional. El análisis de datos toma en cuenta a los 558 postulantes que se registraron en el curso intensivo febrero 2020. Se aplica una técnica de regresión logística binaria, con una variable dependiente dicotómica denominada rendimiento académico y variables dependientes de carácter académico y demográfico. Un hallazgo relevante de esta investigación indica que uno de los principales predictores del rendimiento académico es la modalidad con la que se atiende el curso. Aplicar esta metodología innovadora y tecnológica permite un proceso con mayor tasa de ingreso y de mejor rendimiento académico en comparación con la metodología tradicional.
Palabras Claves análisis logístico binario, proceso de aprendizaje, rendimiento académico.

1 Introducción

Los constantes avances en el área de la tecnología educativa, la necesidad de manejar y analizar gran cantidad de datos y la creciente demanda de profesionales que enfrenten este reto, conduce al replanteamiento de seguir utilizando la modalidad tradicional de enseñanza-aprendizaje o reemplazarla por modalidades innovadoras que centren su proceso en el estudiante y que consten de actividades grupales e individuales, que con el adecuado soporte tecnológico lo hagan responsable de su propia evolución de una manera integral, con el objetivo de que adquiera competencias, habilidades y actitudes que van más allá de la memorización, y así logre un desarrollo adecuado de su carrera universitaria y vida profesional.

Los estudiantes que terminan el nivel secundario y atienden el curso de nivelación y admisión en la ESPOL y en general en toda institución de educación superior en Ecuador, adolecen de conocimientos sólidos de los contenidos de las materias de ciencias exactas y de hábitos de estudio (Álvarez, Baquerizo, Noboa, García-Bustos, y Mera, 2020); si lo anterior se suma a la reconocida rigurosidad académica de la ESPOL, el resultado es una baja tasa de ingreso, aún cuando la institución colabora con actividades adicionales al curso tales como tutorías académicas y clases de refuerzo. Salvo pequeñas variaciones porcentuales, la tasa bruta de matriculación en educación superior viene disminuyendo desde el año 2015 (Rosales, 2020).

Estos antecedentes motivaron a la institución a instaurar desde el año 2018 la modalidad de Aprendizaje Activo, como alternativa a la siempre utilizada metodología tradicional, considerando una serie de actividades grupales e individuales basadas

en tecnología, donde el docente es el guía del proceso y el estudiante logra expresar y justificar respuestas de manera escrita y oral, con un sistema de evaluación y asistencia en línea que comprende la utilización de videos, controles de lectura, preguntas conceptuales y procedimentales, ejercicios de aplicación acerca de la temática a tratar en la clase presencial, los cuales son el punto inicial para mejorar el rendimiento académico, logrando un aprendizaje significativo en reemplazo de la consabida memorización. La base conceptual en que se fundamentan los elementos del proceso innovador y su rigurosidad conducen a los estudiantes a la oportunidad y necesidad de desarrollar habilidades, manejar sintaxis y semántica en el uso de las definiciones en ciencias exactas y a la utilización y comprensión de modelos científicos en la solución de problemas (Álvarez et al., 2020).

El objetivo de esta investigación fue pronosticar por medio de herramientas estadísticas descriptivas e inferenciales y una técnica de regresión logística binaria, el rendimiento académico de los postulantes del área de ciencias exactas en el sistema de admisión de la ESPOL, considerando el grupo de los estudiantes que atendió el curso con Aprendizaje Activo, así como los que lo hicieron de forma tradicional.

2 La metodología de Aprendizaje Activo

Luego de un proceso de selección que involucra variables de naturaleza académica, social y demográfica se constituyen paralelos de 50 estudiantes (10 grupos de 5 estudiantes) guiados por tres docentes. Para cada una de las asignaturas de ciencias exactas que se imparten en el curso de admisión se desarrollan 7 actividades dentro de cada capítulo de contenido del curso.

La figura A1 (Apéndice A) muestra el ciclo de la metodología de Aprendizaje Activo y las actividades que la componen.

La primera actividad es de tipo autónoma, en la cual el estudiante a través de videos y una guía instruccional de lectura que detalla el tema a tratar, los objetivos, ejemplos y páginas a revisar en el libro guía, prepara el contenido de la clase. Al inicio de la clase el estudiante desarrolla un control de lectura en dos rondas: individual y grupal, a través de una plataforma online. Luego, por medio de una retroalimentación se proyectan los resultados y se absuelven dudas de cada pregunta. A continuación, se desarrolla la actividad del taller que consiste en resolver de manera grupal 2 o 3 ejercicios, pudiendo realizarse consultas que no sean digitales; una vez terminado el taller se procede a la respectiva retroalimentación de este. Con el control de lectura y el taller finaliza la clase.

Estas actividades se realizan de manera continua hasta terminar el contenido de un capítulo y luego se desarrolla la clase vía streaming como siguiente actividad, donde profesores y estudiantes fuera del horario de clase se conectan por medio de una sesión virtual y disipan dudas sobre los ejercicios de la tarea que estructuran el banco a ser utilizado en la actividad de exposición de ejercicios.

En la siguiente clase se desarrolla la exposición de la tarea, en la cual se asigna un ejercicio a cada grupo, el mismo que debe ser resuelto en la pizarra de cada gru-

po, de manera ordenada y justificada, teniendo además que responder las preguntas realizadas por los profesores, además, mientras un grupo expone, el resto de los estudiantes deben permanecer atentos a la actividad.

La siguiente actividad es el tutorial, donde los estudiantes de manera grupal y con límite de tiempo deben resolver problemas de nivel crítico con acceso a soporte de consulta de cualquier tipo, una vez resuelto el ejercicio deben subir la foto del desarrollo de este a la plataforma en línea para la revisión y calificación respectiva.

Finalmente se desarrolla la actividad de mayor ponderación, denominada prueba de salida, que se realiza de manera individual y grupal y abarca todo el contenido de la unidad respectiva.

Las tablas B1 B2 (Apéndice B) detallan la ponderación de las actividades que se realizaron en las modalidades de aprendizaje activo y tradicional, respectivamente, en el curso de admisión en referencia.

3 Indicadores del rendimiento académico

Tradicionalmente el rendimiento académico se ha caracterizado de manera simplista al ser relacionado a las notas promedios de los estudiantes cuando se evalúa un determinado contenido, especialmente en las instituciones de educación superior el rendimiento académico es complicado de identificar y definir, debido a sus características. Fullana (1996) enfatiza en el foco multidimensional del rendimiento académico y recalca que es el fin del proceso académico en el cual coinciden los efectos de muchas variables sociales, individuales, académicas y de todas sus relaciones. Rodríguez, Fita, y Torrado (2004) recomiendan diferenciar entre el rendimiento inmediato asociado a las notas y el rendimiento mediato asociado a logros profesionales y personales.

Investigar las formas de mejorar el rendimiento académico, los factores que lo determinan, las variables influyentes, las metodologías de aprendizaje más apropiadas, proviene de muchos atrás ya que su análisis es relevante para mejorar la calidad académica. Es así, que existe una cantidad significativa de estudio teóricos y empíricos en el entorno preuniversitario y universitario (Garnica, González, Díaz, y Torres, 1991; P. González, 1982; Diaz, 1995; Roselli, 2008; McArdle, Paskus, y Boker, 2013; Cardona, Vélez, y Tobón, 2016; Padua, 2019).

En lo que respecta a los factores asociados al rendimiento académico se consideran como más influyentes los de tipo pedagógico, institucional, sociodemográfico y psicosocial, todas complejas por sí mismas (Tournon, 1984).

Cuando se pronostica el rendimiento académico se condiciona estimar la variable dependiente en función de variables predictoras de tal forma que se pueda evidenciar el éxito o fracaso del estudiante en un escenario específico. Sostener que el mejor predictor del rendimiento académico futuro es el rendimiento previo ha sido el resultado de la investigación de varios autores (De Miguel y Arias, 1999; Rodríguez Ayán, 2007; Tomás, Expósito, y Sempere, 2014).

Roselli (2008) desarrolla un trabajo cuyo objetivo es la comparación entre dos modelos de enseñanza en la universidad, con la disyuntiva individual-grupal; el individual centrado en el sistema tradicional de aprendizaje y el grupal centrado en que el sujeto de enseñanza-aprendizaje lo forman grupos de 4 estudiantes que asistían a sesiones de teoría y práctica, trabajando colaborativamente pero incluyendo la supervisión del profesor, con antelación se suministraba la bibliografía que producía los diversos trabajos parciales entregados por grupo. Una conclusión importante de este trabajo es que el aprendizaje grupal funciona mejor con los alumnos más aprovechados académicamente, al parecer la autorregulación requiere como prerrequisito un determinado nivel de capacidades iniciales.

También son numerosos los estudios en los que se ha utilizado la técnica de regresión logística binaria por medio de herramientas estadísticas para predecir el rendimiento académico (Valera, Sinha, Varela, y Ponsot, 2009; Ibarra y Michalus, 2010; Heredia y Calderón, 2014; Vitola, 2015).

4 Método, población y variables de estudio

La investigación desarrollada tiene un enfoque retrospectivo, cuantitativo, observacional y de técnica multivariante. El instrumento empleado en la recogida de datos fue el sistema académico de la Dirección de Admisiones de la Escuela Superior Politécnica del Litoral, luego de lo cual se procedió al análisis y depuración del conjunto de datos proporcionado, tanto para la estadística descriptiva, contraste de hipótesis y análisis de regresión logística binaria; técnicas estadísticas que fueron ejecutadas con el software estadístico R versión 1.3.1056 para Mac OS.

La muestra de estudio incluye a 588 postulantes registrados en el curso de admisión intensivo 2020 en el área de ciencias e ingeniería, 258 de ellos incluidos en la modalidad de aprendizaje activo repartidos en 6 paralelos y 300 postulantes en la modalidad tradicional repartidos en 8 paralelos.

Son requisitos para que el postulante sea considerado en el criterio de inclusión en la población de estudio que haya concluido sus estudios de nivel secundario, haber realizado el examen de acceso a la educación superior, responsabilidad de la Secretaría de Educación, Ciencia, Tecnología e Innovación del gobierno nacional y haber realizado el examen de admisión ESPOL habiendo obtenido una calificación mayor o igual a 40 puntos y menor que 60 puntos.

La tabla C1 (Apéndice C) muestra datos generales de sexo, número de postulantes, número de aprobados, número de repetidores de curso de la población de estudio, dependiendo de la modalidad de estudio.

En lo referente a las variables de estudio, se considera como variable dependiente tipo dicotómica al rendimiento académico (0= reprueba, 1=aprueba) y se dispone de 7 variables independientes, 4 de ellas de tipo cuantitativa y 3 de tipo cualitativa. En general, el conjunto de datos consta de 558 observaciones y 8 variables, cuya estructura en R es la siguiente: "*mod*" (modalidad), "*edad*"(edad), "*sexo*"(sexo), "*rep*"(repetidor), "*mat*"(nota del examen de admisión de matemáticas), "fis"(nota

del examen de admisión de física), "*paa*"(nota de la prueba de aptitud académica), "*racad*"(rendimiento académico).

Fueron dicotomizadas las variables explicativas *modalidad* (1 = aprendizaje activo, 0 = tradicional), *sexo* (0 = masculino, 1 = femenino) y *repetidor* (0 = no repetidor, 1 = repetidor), tomando en cuenta la revisión bibliográfica previa y lo recomendado por Carballo y Guelmes (2016).

5 Metodología

Un modelo de regresión logística permite determinar la probabilidad p de que ocurra un evento A dependiendo de los valores de ciertas variables $X_1, \ldots, X_p$, es decir, si $x = (x_1, \ldots, x_p)'$ son las observaciones de un individuo sobre las variables, entonces la probabilidad $p(x)$ de que acontezca A es $p(y = 1/x)$, siendo $p(y = 0/x) = 1 - p(x)$ la probabilidad de que A no suceda dado x, debiendo notar que al estar $p(x)$ comprendido entre 0 y 1 no podemos asumir que $p(x)$ sea una función lineal. Entonces es conveniente suponer un modelo lineal para la denominada transformación logística de la probabilidad.

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + \beta' x \tag{1}$$

siendo $\beta = (\beta_1, \ldots, \beta_p)'$ los parámetros de regresión. El modelo (1) equivale a suponer las siguientes probabilidades para el evento y su contrario, ambas en función de x:

$$p(x) = \frac{e^{\beta_0+\beta'_x}}{1+e^{\beta_0+\beta'_x}}; \qquad 1 - p(x) = \frac{1}{1+e^{\beta_0+\beta'_x}} \tag{2}$$

Si comparamos con el modelo de regresión lineal:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e \tag{3}$$

podemos comprender el modelo logístico en el sentido de que $y = p(x)+e$, donde y solo toma valores 0 ó 1, es decir, si $y = 1$ entonces $e = 1 - p(x)$ con probabilidad $p(x)$ y si $y = 0$ entonces $e = -p(x)$ con probabilidad $1 - p(x)$. Así, dado x, el error e tendrá media 0 y varianza $p(x)(1 - p(x))$.

Dada una observación del evento, la regla de discriminación logística simplemente decidirá si un individuo posee la característica del evento si $p(x) > 0{,}5$ y no la posee si $p(x) \leq 0,5$, entonces si introducimos la función discriminante:

$$L_g(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] \tag{4}$$

la regla de decisión logística es: si $L_g(x) > 0 \rightarrow y = 1$, si $L_g(x) \leq 0 \rightarrow y = 0$ (Cuadras, 2019).

Una vez obtenido el modelo lineal generalizado (incluidas todas las variables explicativas) se seleccionaron las variables dependientes estadísticamente signifi-

cativas, es decir, aquellas que a un nivel de significancia de $\propto = 0{,}05$ su valor de estadístico de Wald es en valor absoluto mayor al punto crítico $Z_{\propto/2} = 1{,}96$, tal como lo señala Cañadas (2013); lo cual nos conduce al modelo ajustado.

Al realizar el contraste de los parámetros del modelo ajustado se obtuvo el estadístico de prueba que tiene una distribución chi-cuadrada con d grados de libertad y una vez planteadas las hipótesis:

$$H_0 : \text{el modelo no es significativo } (\beta_1 = \beta_2 = \cdots = \beta_r = 0)$$
$$H_a : \text{el modelo sí es significativo}$$

si el valor p del estadístico de prueba, calculado utilizando el lado derecho de la distribución (región de rechazo), es menor al valor de significancia, se rechaza la hipótesis nula y el modelo será significativo.

Luego se procedió a calcular intervalos de confianza con la prueba de Wald, el cual se basa en que los parámetros β_r siguen de manera asintótica una distribución normal $N(\beta_r, \hat{\sigma}^2(\hat{\beta}_r))$:

$$p\left[-Z_{\propto/2} \leq \frac{\hat{\beta}_r - \beta_r}{\hat{\sigma}(\hat{\beta}_r)} \leq Z_{\propto/2}\right] = 1 \tag{5}$$

con lo que el intervalo aproximado para el parámetro β_r a un nivel $(1- \propto)$ es $\hat{\beta}_r \pm Z_{\propto/2}\hat{\sigma}(\hat{\beta_r})$ si el intervalo de confianza de los exponenciales asociado a alguna variable explicativa incluye al 1 no se puede rechazar la hipótesis nula de que $\beta_r = 0$ al nivel de significancia elegido.

Los valores estimados por el modelo se calculan directamente en R y se almacenan en el objeto "glm", se puede acceder a ellos con el operador $ o con la función "fitted.values".

Para evaluar la adecuación global del modelo al conjunto de datos y detectar la presencia de valores influyentes y anómalos, realizar el análisis de residuos es de vital importancia, en este estudio se utilizaron los de Pearson y de la devianza.

Se calculó el estadístico de Hosmer-Lemeshow que opera una partición de datos con base en probabilidades predichas para luego analizar la tabla de contingencia a través del estadístico X^2 asociado, ya que no existe garantía alguna en la práctica de que un modelo de regresión logística se encuentre adecuadamente ajustado a los datos. Para las medidas tipo R^2 se calcularon las de McFadden, la de Cox&Snell y Nagelkerke, también se predijo el poder de predicción del modelo por medio de la tasa de clasificaciones correctas y el análisis de curvas ROC (receiver operating characteristics).

Con el fin de diagnosticar y validar el modelo se analizó la forma en que las observaciones afectan al modelo, utilizando el análisis de residuos y los valores influyentes; obtener los valores de las medidas de influencia permitió ubicar los valores influyentes analizando la forma en que afectan a los parámetros del modelo, esto se logró calculando las distancias de Cook.

Por medio del factor de inflación de la varianza generalizado (GVIF) se determinó la posible multicolinealidad entre las variables, si este valor se encuentra próximo a 1, una de las variables está muy poco relacionada con la otras, es decir, ausencia de colinealidad (Fox y Monnette, 1992).

Uno de los problemas mas repetitivos es que el modelo no se ajusta bien a otro conjunto de datos que no sea el del estudio, es decir, que no sea generalizable, este problema denominado sobreajuste se detecta por medio de la validación cruzada, utilizando la función k-Fold, donde la muestra se divide en k sub-muestras para utilizar $k-1$ de ellas y estimar el modelo y las demás como submuestras de evaluación, el proceso es repetido k veces, haciendo que cada sub-muestra se use una vez para la evaluación del modelo y $k-1$ veces para el respectivo ajuste.

6 Resultados

6.1 Análisis estadístico univariante del conjunto de datos

La tabla 1 muestra los datos descriptivos más importantes de las 4 variables cuantitativas.

Tabla 1: Resumen descriptivo – variables cuantitativas

Variable	min	1Q	med	prom	3Q	max	moda	SD	Asim.	Kurt.
edad	17	18	19	18,8	19	32	18	1,41	4,81	39,2
mat	0	21	32	32,8	43	100	40	17,2	1,12	5,93
fis	0	28	41	40,5	51	100	46	16,3	0,39	0,70
paa	49,8	73,8	80	80,8	89	100	100	11,0	−0,11	2,58

Fuente: Elaboración propia

Es relevante destacar los siguientes resultados:

- Significativa dispersión de las notas de los exámenes de admisión de matemáticas y física con respecto a sus respectivas medias.
- Las deficientes notas promedio de los exámenes de admisión de matemáticas y física, siendo 40 la nota que más se repite en el de matemáticas y 46 en el de física.
- Evidente desviación de la normalidad en los datos de las variables *edad*, *mat* y fis con asimetría positiva y distribución leptocúrtica, es decir, con destacable número de datos alrededor del valor central de cada variable.
- Ligera desviación de la normalidad en los datos de la variable *paa* con asimetría negativa y distribución leptocúrtica, es decir, con destacable número de datos alrededor del valor central de cada variable.
- Correlación positiva y media entre las variables *mat*, fis y *paa*.
- Correlación negativa y baja entre la variable *edad*, y las variables *mat*, fis y *paa*.

	edad	mat	fis	paa
edad	1,00	−0,20	−0,24	−0,26
mat	−0,20	1,00	0,40	0,47
fis	−0,24	0,40	1,00	0,49
paa	−0,26	0,47	0,49	1,00

Las tablas D1 y D2 (Apéndice D) muestran la información de número y porcentaje de postulantes aprobados (AP), reprobados (RP), no repetidores aprobados (No REP AP), no repetidores reprobados (no REP RP), repetidores aprobados (REP AP) y repetidores reprobados (REP RP) para ambas modalidades.

Es importante resaltar que aprobaron el curso 69 % (178 postulantes) que siguieron aprendizaje activo y solo 43,7 % (131 postulantes) en modalidad tradicional.

Los datos de las variables cuantitativas no provienen de una distribución normal ya que el p-value obtenido por medio de la prueba Jaque Bera es menor al valor de significancia y se rechaza la hipótesis nula de que los datos provienen de distribución normal. En la figura 1 se muestra el histograma con curva normal teórica para la variable fis.

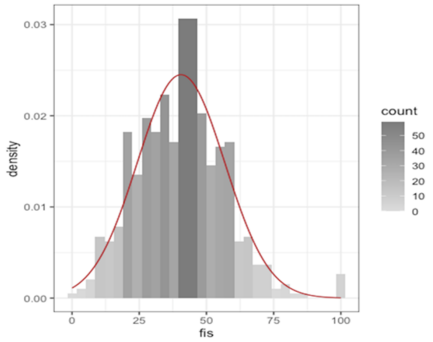

Figura 1: Histograma con curva normal teórica para fis
Fuente: Elaboración propia en software R

Entre las variables cualitativas se realizó la prueba de independencia, resultando que las variables *sexo* y *mod* no están relacionadas y sí lo están las variables *mod* y *rep*, *sexo* y *rep*. Además, al aplicar la prueba chi-cuadrado para outliers se encontró que todas las variables cualitativas tienen presencia de valores outliers, *edad* (49), *mat* y fis (100), *paa* (49,8).

En lo que se refiere a la homogeneidad de las varianzas o supuesto de homocedasticidad, al aplicar la prueba de Levene a las variables *paa* y *mat* no se encontró evidencia suficiente para considerar que existe diferencia significativa entre las varianzas de las variables que representan las notas promedio del examen de admisión de matemáticas y de la prueba de aptitud, por grupos de modalidad de aprendizaje, debido a que en ambos casos se obtuvo un p-valor mayor al valor de significancia; sin embargo para la variable fis al aplicar la misma prueba se obtuvo un p-valor de

0,0004 evidenciando diferencia significativa en las varianzas de la variable que representa la nota promedio del examen de física por modalidad de estudio. El tamaño de efecto respectivo para la variable fis medido con el índice de Cohen, resultó ser alto (0,99). En la figura 2 se puede observar el gráfico comparativo de medias poblacionales por modalidad de estudio para la variable *mat*.

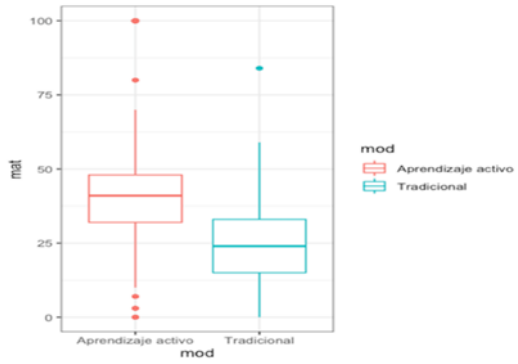

Figura 2: Comparación de medias poblacionales para fis por modalidad
Fuente: Elaboración propia en software R

6.2 Regresión logística binaria

En el modelo generalizado la variable *edad* resultó ser estadísticamente no significativa al analizar su valor de significancia (0,535) y del estadístico de Wald (0,384) ya que este último estadístico es en valor absoluto menor al punto crítico $Z_{(\propto/2)} = 1,96$, haciendo que se rechace la hipótesis nula de que los coeficientes sean iguales a cero.

Se realizó posteriormente el proceso de selección de variables para obtener el modelo ajustado, donde no aparece la variable *edad*, el modelo ajustado es mejor que el modelo nulo (solo intercepto) y el modelo generalizado teniendo los mejores índices de ajuste (estadístico de prueba con distribución chi-cuadrada y valor de 120,5 con 5 grados de libertad y p-valor nulo), un AIC (coeficiente de información de Akaike) de 660,6, BIC (coeficiente de información bayesiano) de 690,9 y devianza residual de 646,6.

Con respecto a los coeficientes, el efecto predominante en la explicación del modelo le pertenece a la variable *mod* de tal forma que las probabilidades de que apruebe el curso un postulante que sigue aprendizaje activo son $2,14$ $(1/0,468)$ veces superiores a las de quien lo hace con metodología tradicional. Por cada punto adicional que se obtenga para la variable *mat* la probabilidad de aprobar es $1,02$ veces mayor; $1,05$ veces mayor para la variable fis y $1,04$ veces mayor para la

variable *paa*. En los intervalos de confianza se puede notar que el número 1 no está incluido en ninguno de ellos, por lo que se ratifica que todas las variables consideradas son estadísticamente significativas.

La tabla 2 muestra la estimación de los efectos fijos del modelo ajustado incluyendo los coeficientes, errores estándar, oods-ratio o exponencial de los parámetros y el intervalo de confianza al 95 % para los oods-ratio.

Tabla 2: Resumen del modelo ajustado

	Coeficiente regresión (B)	Error estándar	Exp ODDS (B)	I.C. al 95 % para Exp.(B)	
				Inferior	Superior
Intercepto	−4,22	0,896	0,004		
$mod(1)$	−0,76	0,283	0,468	0,27	0,81
$sexo(1)$	−0,58	0,220	0,561	0,36	0,86
$rep(1)$	−1,30	0,363	0,273	0,13	0,55
mat	0,02	0,008	1,022	1,01	1,04
fis	0,05	0,009	1,054	1,04	1,07
paa	0,04	0,012	1,040	1,02	1,06

Fuente: Elaboración propia

Los valores predichos de las 8 primeras observaciones, que indican la probabilidad predicha de aprobación del postulante y que deben ser comparadas con el rendimiento académico son:

1	2	3	4	5	6	7	8
0,829	0,872	0,826	0,831	0,495	0,764	0,667	0,472

Los valores para las medidas tipo R^2 que resultaron son: McFadden (0,16), Nagelkerke (0,26) y Cox&Snell (0,20).

Utilizando la función respectiva de la lista de correos R-help, modificándola para que realice el contraste de hipótesis para el cálculo de los coeficientes de Hosmer-Lemeshow y aplicando la función "homerslem" al modelo con el fin de repartir los datos en base a los cuantiles de la distribución (grupos más homogéneos), se obtiene un valor de X-squared de 3,6 y un p-valor de 0,89 con lo que se evidencia que el modelo ajustado se ajusta globalmente a los datos.

En referencia a que existen 309 postulantes aprobados y 249 reprobados, el modelo predijo que 326 aprueban y 232 reprueban con la tabla de clasificación, la cual nos indica que 149 postulantes que no aprueban, el modelo predice que no aprueban y que 226 que, sí aprueban, el modelo predice que sí aprueban.

	predicción	
	0	1
0	149	100
1	83	226

Relacionado al punto de corte calculado (0,605), la tasa de clasificaciones correctas para el total de los postulantes del estudio es de 67,2 %. En la figura 3 se muestra la tasa de clasificaciones correctas para diversos puntos de corte y para el punto de corte calculado en el estudio.

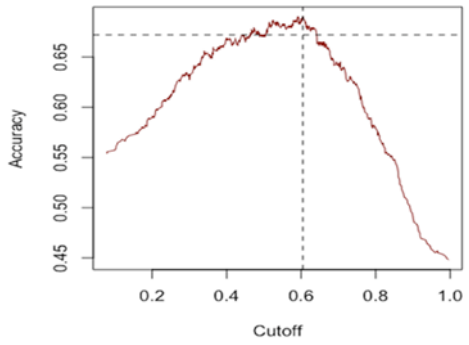

Figura 3: Tasa de clasificaciones para diversos puntos de corte
Fuente: Elaboración propia en software R

Se puede observar en la figura 4 que se presenta la curva ROC del modelo, se puede observar adicionalmente que el área bajo ella es de es de 0,75 unidades cuadradas.

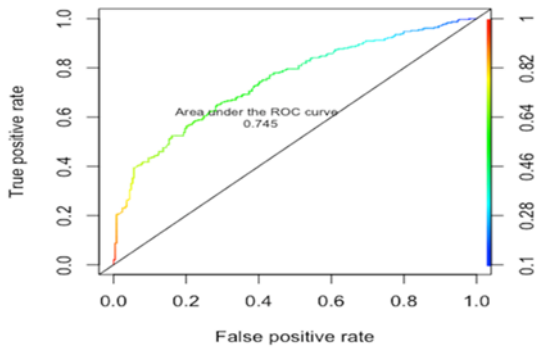

Figura 4: Curva ROC del modelo
Fuente: Elaboración propia en software R

Una vez calculados los residuos de Pearson y de la devianza estandarizados se comprobó que sólo 2 de ellos resultan ser significativos ya que son en valor absoluto mayores a 2, como lo muestra la figura 5.

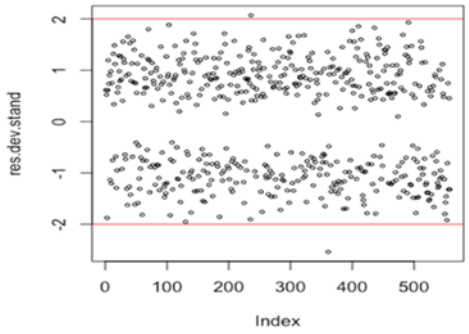

Figura 5: Residuos de la devianza estandarizados
Fuente: Elaboración propia en software R

Al calcular las medidas de influencia se determinaron las distancias de Cook respectivas y no se encontraron distancias mayores a la unidad, por lo tanto, no existen valores influyentes que afecten las estimaciones del modelo (Cook y Weisberg, 1982). Tampoco se encontraron valores elevados de factores generalizados de inflación de la varianza, es decir, no existe colinealidad.

En la validación cruzada, al comprobar si el modelo se ajusta bien a un conjunto de datos distinto al utilizado en este estudio, el valor resultante de tasa media de clasificaciones incorrectas es 33,3 % y se evidencia que en media, el modelo clasifica correctamente a más del 66,5 % de los postulantes cuando sus propios datos no han sido utilizados en el ajuste del modelo.

La expresión lineal del modelo ajustado es:

$$logit(racad) = -4{,}22 - 0{,}76\, mod - 0{,}58\, sexo - 1{,}3\, rep + 0{,}022\, mat + 0{,}053\, fis + 0{,}04\, paa \tag{6}$$

y su probabilidad condicionada queda determinada por la expresión:

$$p(aprobar) = \frac{e^{-4{,}22 - 0{,}76\, mod - 0{,}58\, sexo - 1{,}3\, rep + 0{,}022\, mat + 0{,}053\, fis + 0{,}04\, paa}}{1 + e^{-4{,}22 - 0{,}76\, mod - 0{,}58\, sexo - 1{,}3\, rep + 0{,}022\, mat + 0{,}053\, fis + 0{,}04\, paa}} \tag{7}$$

7 Discusión

Los resultados obtenidos del análisis estadístico descriptivo e inferencial del conjunto de datos ratifican el bajo rendimiento académico que evidencian los postulantes al ingresar a ESPOL en los exámenes de ingreso, siendo Matemáticas la asignatura con mayor deficiencia. A pesar de que se mantiene la tradicional supremacía de la presencia de postulantes de sexo masculino sobre el femenino en el área de ciencias exactas sin importar la metodología de enseñanza-aprendizaje utilizada, es

interesante el hallazgo de que en promedio, en modalidad tradicional aprobaron apenas 3,8 % más hombres que mujeres y en aprendizaje activo aprobaron 10,7 % más mujeres que hombres, con el antecedente de que en modalidad tradicional el 61 % de los hombres que aprobaron eran repetidores de curso. No se puede afirmar radicalmente que exista una relación directa entre el sexo y el rendimiento académico pero hay estudios que le dan a la mujer una ligera tendencia a tener mayor rendimiento académico que el hombre (F. A. González, 1996; Lladó, Rodríguez, y Torrado, 2004). Además, en promedio aprobaron 27,3 % más postulantes que atendieron el curso con aprendizaje activo en relación con los que lo hicieron con modalidad tradicional, hallazgo que ratifica el único análisis global que se había realizado con la información hasta que se realizó esta investigación.

Un hallazgo muy preocupante que entrega esta parte de la investigación es que solo el 42 % de los postulantes repetidores, prácticamente todos ellos en modalidad tradicional, lograron aprobar el curso a pesar de que repetían los mismos contenidos y la experiencia de ya haber tomado el curso.

A pesar de que en ninguno de los casos existe una correlación significativa entre las variables explicativas del estudio, es importante señalar que a mayor edad del postulante disminuyen en menor grado las notas obtenidas en los exámenes de ingreso de Matemáticas, Física y en la prueba de aptitud académica, y que están correlacionadas de manera positiva y media las variables *mat*, *paa* y fis, indicando que los estudiantes que logran rendimiento medio en uno de los exámenes también lo hacen en los otros dos.

En lo que se refiere a la construcción del modelo generalizado de regresión logística binaria para predecir el rendimiento académico de los postulantes, la variable *edad* resultó ser estadísticamente no significativa, lo cual fue corroborado cuando se construyó el modelo ajustado, el mismo que fue el más adecuado, por ser el modelo más reducido que explicó los datos (principio de parsimonia) y además es técnicamente congruente e interpretable, con una tasa de clasificación correcta bastante aceptable.

A pesar de que se obtuvo valores bajos de las medidas tipo R^2, es importante recalcar que si bien es posible que predictores adicionales puedan incrementar la potencia explicativa del modelo también es posible que los datos contengan una cantidad inherentemente más alta de inexplicable variabilidad; en todo caso, aún cuando las medidas tipo R^2 son bajas, los p-valores obtenidos indican una relación real entre los predictores significativos y la variable respuesta (Visbal, 2019).

Se considera que un modelo es mejor que otro si la curva ROC se acerca al borde superior izquierdo, o lo que es lo mismo, que el área bajo la curva sea mayor (Franco-Nicolás y Vivo-Molina, 2007) y nuestro modelo cumple satisfactoriamente con esta consideración.

Este trabajo también consideró el hecho de analizar residuos con el fin de aislar los datos en los cuales el modelo se ajustaba mal, así como de los que ejercían una influencia excesiva sobre el mismo. También se pudo comprobar la ausencia de valores elevados de factores generalizados de inflación de la varianza, es decir, no existe colinealidad. La validación cruzada demostró que el modelo se ajusta bien a conjuntos de datos distintos a los utilizados en este estudio.

Con base en el análisis de los coeficientes del modelo, quedó evidenciada la predominancia y ventajas de la aplicación del modelo de Aprendizaje Activo en el sistema de admisión de ESPOL con respecto al modelo tradicional y la incidencia del estado de repetidor de curso del postulante.

A manera de posición personal ante los hallazgos de esta investigación, es pertinente considerar que se debe profundizar en el uso de técnicas estadísticas y el estudio de variables adicionales, que permitan evidenciar la efectividad de un modelo alternativo e innovador de enseñanza-aprendizaje sobre el modelo tradicional. Se espera que esto sirva de sustento y aporte académico en el momento de generar políticas administrativas y económicas que ayuden a mejorar el rendimiento académico de los bachilleres y al desarrollo de habilidades cognitivas y actitudinales. La finalidad es el desenvolvimiento adecuado del estudiante universitario, incrementar las tasas de ingreso, reducir las tasas de deserción, y adicionalmente, justificar científicamente que el presupuesto asignado para la implementación de este tipo de modalidades centradas en el estudiante siempre será una inversión con un resultado favorable y prometedor para la sociedad.

8 Conclusiones

En lo que corresponde al análisis descriptivo e inferencial del conjunto de datos, este estudio demuestra que se cumplen todos los supuestos para la aplicación del modelo logit: las variables predictoras son categóricas o continuas, no se requiere linealidad, existe independencia del error y no hay presencia de multicolinealidad.

Resultan interesantes los hallazgos de que, en general, y contrariamente a lo que se pueda suponer, sólo el 41,7 % de los repetidores pudieron aprobar el curso de admisión; es evidente el bajo rendimiento académico en los exámenes de ingreso y la alta dispersión de esas notas con respecto a su media, lo cual ratifica la poca efectividad del sistema educativo del nivel secundario; los datos de las notas de los exámenes de ingreso poseen desviación de la normalidad con asimetría positiva y distribución leptocúrtica con un considerable número de datos alrededor del valor central de la variable.

La *edad* del postulante a ingresar a la ESPOL no incide en el rendimiento académico, lo cual se ratificó cuando no fue incluida en el modelo ajustado, sin importar el proceso de selección de variables utilizado.

En el modelo ajustado, la variable *mod* es predominante, de tal forma que las probabilidades de que un postulante en aprendizaje activo apruebe el curso de admisión son 2,14 veces superiores a las de quien lo hace con metodología tradicional. El modelo ajustado posee mejores índices de ajuste que los modelos nulo y generalizado, es significativo y se ajusta globalmente a los datos, posee una tasa de clasificaciones correctas del 67,2 %. Sólo existen 2 residuos de devianza significativos que resultan despreciables en relación con el total (menos del 0,36 %); valores predichos cercanos a cero se corresponden mayoritariamente con valores observados iguales a cero y viceversa; en nuestro estudio no existen distancias de Cook significativas que pro-

voquen la presencia de valores influyentes que puedan influir en las estimaciones del modelo. Con base en la tasa de clasificaciones correctas calculada mediante el proceso de validación cruzada, la ausencia de valores influyentes y la mínima cantidad de residuos significativos, se puede concluir que el modelo ajustado no adolece de falta de ajuste ni de problemas de sobreajuste.

En virtud de los resultados de esta investigación, se pueden inferir 2 importantes implicaciones de interés para la educación pública de nivel superior en Ecuador: la posibilidad de incrementar el rendimiento académico y la tasa de ingreso de los postulantes a ingresar a la ESPOL y el rol predominante de la modalidad de Aprendizaje Activo en el sistema de admisión y nivelación de la referida institución de educación superior; así, esta investigación puede servir de base para replantear las políticas de admisión en las demás universidades públicas del país.

Finalmente, como acción prospectiva de investigación, es recomendable estudiar el efecto de la utilización de la metodología innovadora de Aprendizaje Activo en el desempeño académico del estudiante durante su carrera universitaria, inicialmente al término del primer año de la carrera, y en función de los recursos humanos y económicos que demanda la implementación de esta modalidad alternativa de enseñanza-aprendizaje, aplicarla en materias fundamentales de la malla curricular respectiva, así como también estudiar el estilo de aprendizaje de los postulantes con el fin de establecer académicamente sus preferencias de aprendizaje ante un modelo tradicional o centrado en el estudiante, situación que podría mejorar significativamente la forma en que actualmente se escogen a los postulantes para cada modalidad.

9 Agradecimientos

Se agradece al Mg. Dalton Noboa Macías, ex Director de nivelación y admisiones de la Escuela Superior Politécnica del Litoral por su autorización y apoyo en el proceso de obtención de los datos que permitieron seleccionar, organizar y depurar la información para esta investigación, así como su permanente interés en la continuidad de esta.

10 Bibliografía

Referencias

Álvarez, I., Baquerizo, G., Noboa, D., García-Bustos, S., y Mera, E. (2020). Una nueva metodología de clase invertida aplicada como un programa piloto a estudiantes aspirantes a ingresar en una universidad ecuatoriana. *18th LACCEI*

International Multi-Conference for Engineering, Education and Technology, 1–7. doi: 10.18687/LACCEI2020.1.1547

Cañadas, J. (2013). *Regresión logística. Tratamiento computacional con R.* Universidad de Granada. Granada, España. doi: 10.13140/RG.2.1.1342.6083

Carballo, M., y Guelmes, E. (2016). Algunas consideraciones acerca de las variables en las investigaciones que se desarrollan en educación. *Revista Universidad y Sociedad*, *8*(1), 140–150.

Cardona, S., Vélez, J., y Tobón, S. (2016). Contribución de la evaluación socioformativa al rendimiento académico en pregrado. *Educar*, *52*(2), 423–447.

Cook, D., y Weisberg, S. (1982). *Residuals and Influence in Regression*. (N. Y. C. and Hall (ed.)). Descargado de `https://hdl.handle.net/11299/37076`

Cuadras, C. (2019). *Nuevos métodos de análisis multivariante*. (CMC Editions).

De Miguel, M., y Arias, J. (1999). La evaluación del rendimiento inmediato en la enseñanza universitaria. *Revista de Educación*, *320*, 356.

Diaz, F. (1995). La predicción del rendimiento académico en la Universidad: un ejemplo de aplicación de la regresión múltiple. *Enseñanza*, *13*, 43–61.

Fox, J., y Monnette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(147), 178–183.

Franco-Nicolás, y Vivo-Molina, J. (2007). *Análisis de Curvas ROC: Principios básicos y aplicaciones.* Madrid: Editorial La Muralla.

Fullana, J. (1996). La investigación sobre las variables relevantes para la prevención del fracaso escolar. *Revista Investigación Educativa*, *14*(1), 63–90.

Garnica, E., González, P., Díaz, A., y Torres, E. (1991). Análisis discriminante: Estudio del rendimiento estudiantil. *Economía*, *16*(6), 51–77.

González, F. A. (1996). Comprensión lectora y rendimiento académico. *Revista Gallega de Psicopedagogía*, *13*(9), 209–221.

González, P. (1982). Análisis estadístico del rendimiento estudiantil en la Universidad de los Andes. *Merida, Venezuela. Facultad de Ciencias, ULA (mimeografía)*.

Heredia, Y., y Calderón, D. (2014). Factores que afectan el desempeño académico. *Ciudad de México Editoras*, 251–258.

Ibarra, M., y Michalus, J. (2010). Análisis del Rendimiento Académico mediante un modelo Logit. *Ingeniería Industrial*, *9*(2), 47–56.

Lladó, E., Rodríguez, S., y Torrado, M. (2004). El rendimiento académico en la transición secundaria-universidad. *Revista de Educación*, *334*, 391–414.

McArdle, J., Paskus, T., y Boker, S. (2013). A Multilevel Multivariate Analysis of Academic Performances in College based on NCAA Student-Athletes. *Multivariate Behavioral Research*, 57–95.

Padua, L. (2019). Factores individuales y familiares asociados al bajo rendimiento académico en estudiantes. *Revista Mexicana de Investigación Educativa*, *24*(80), 173–195.

Rodríguez, S., Fita, E., y Torrado, M. (2004). El rendimiento académico en la transición secundaria - universidad. *Revista de Educación*, *334*, 391–414.

Rodríguez Ayán, M. N. (2007). Análisis multivariado del desempeño académico de estudiantes universitarios de Química. *Universidad Autónoma de Madrid*.

Rosales, B. (2020). *Boletín anual - Senescyt.*

Roselli, N. (2008). La disyuntiva individual-grupal. Comparación entre dos modelos alternativos de eseñanza en la universidad. *Ciencia, Docencia y Tecnología*, *36*(19), 87–118.

Tomás, J., Expósito, M., y Sempere, S. (2014). Determinantes del rendimiento académico en los estudiantes de grado. Un estudio en administración y dirección de empresas. *Revista de Investigación Educativa*, *32*, 379–392.

Tournon, J. (1984). *Factores del rendimiento académico en la universidad.* Ediciones Universidad de Navarra.

Valera, J., Sinha, S., Varela, J., y Ponsot, E. (2009). Una explicación del rendimiento estudiantil universitario mediante modelos de regresión logística. *Visión Gerencial*, *2*, 415–427.

Visbal, D. (2019). *Análisis del rendimiento académico de los estudiantes de la Universidad del Magdalena según variables socioeconómicas y familiares*. Universidad Politécnica de Valencia.

Vitola, L. (2015). Regresión logística: una aplicación en la identificación de variables que inciden en el rendimiento académico, en el área de matemáticas. *Educación y Desarrollo Social*, *1*, 118–1317.

Apéndice A. Ciclo de actividades – modalidad Aprendizaje Activo

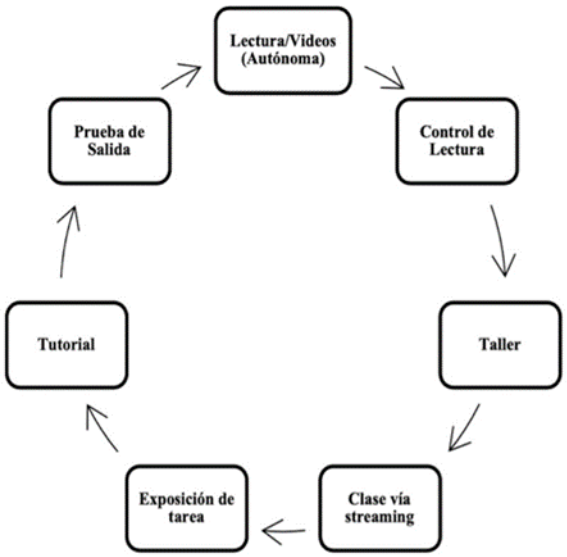

Figura A 1
Fuente: Elaboración propia

Apéndice B. Políticas de evaluación

Tabla B 1: Políticas de evaluación - tradicional

Componentes Mod. Tradicional	Ponderación de la actividad	Ponderación Nota Final	Recuperación
Examen Final	40 %	40 %	El examen de recuperación reemplaza sólo estos componentes
Lección General 1	60 %	35 %	
Lección General 2			
Lección General 3			
Gestión del Aprendizaje: Evaluaciones parciales, Talleres, Trabajo Autónomo.		25 %	Componente no recuperable
Nota Final	100 %	100 %	

Fuente: Elaboración propia

Tabla B 2: Políticas de evaluación - Aprendizaje Activo

Componentes Mod. Aprendizaje Activo	Ponderación de la actividad	Ponderación Nota Final	Recuperación
Examen Final		40 %	El examen de recuperación reemplaza sólo estos componentes
Prueba de Salida		20 %	
Controles de Lectura	20 %	40 %	Componente no recuperable
Talleres	25 %		
Exposición de Tareas	25 %		
Tutoriales	30 %		
Nota Final	100 %	100 %	

Fuente: Elaboración propia

Apéndice C. Datos generales de la población de estudio por modalidad

Tabla C 1

Modalidad	Número Postulantes	Número Repetidores	Sexo		Número Aprobados
			M	F	
Tradicional	300	234	206	94	131
Aprendizaje Activo	258	13	196	62	178

Fuente: Elaboración propia

Apéndice D. Datos de aprobados y reprobados

Tabla D 1: Datos de aprobados y reprobados – Aprendizaje Activo

Sexo	AP	RP	No REP AP	No REP RP	REP AP	REP RP
Masculino	(137)	(59)	(137)	(51)	(0)	(8)
	63,4 %	73,7 %	77,4 %	75 %	0 %	100 %
Femenino	(41)	(21)	(40)	(17)	(1)	0
	30,6 %	26,3 %	22,6 %	25 %	100 %	0 %

Fuente: Elaboración propia

Tabla D 2: Datos de aprobados y reprobados – tradicional

Sexo	AP	RP	No REP AP	No REP RP	REP AP	REP RP
Masculino	(83)	(123)	(21)	(27)	(62)	(96)
	63,4 %	72,8 %	72,4 %	73 %	60,8 %	72,7 %
Femenino	(48)	(46)	(8)	(10)	(40)	(36)
	36,6 %	27,2 %	27,6 %	27 %	39,2 %	27,3 %

Fuente: Elaboración propia

Notación:

- AP: aprobados
- RP: reprobados
- No REP AP: no repetidores, aprobados
- No REP RP: no repetidores, reprobados
- REP AP: repetidores aprobados
- REP RP: repetidores reprobados