Fecha de presentación: diciembre, 2019, **Fecha de Aceptación**: enero, 2020, **Fecha de publicación**: marzo, 2020

18

MODELO LINEAL MIXTO GENERALIZADO PARA LA DETECCIÓN DE ESTUDIANTES EXITOSOS DEL PRIMER AÑO DE INGENIERÍA DE SISTEMAS

GENERALIZED MIXED LINEAR MODEL FOR THE DETECTION OF SUCCESSFUL FIRST-YEAR SYSTEM ENGINEERING STUDENTS

Efraín Díaz Macías[1]
E-mail: efraindiaz@uteq.edu.ec
ORCID: https://orcid.org/0000-0003-4087-029X
Jorge Guanin Fajardo[1]
E-mail: jorgeguanin@uteq.edu.ec
ORCID: http://orcid.org/0000-0001-9150-4009
[1] Universidad Técnica Estatal de Quevedo. Ecuador.

Cita sugerida (APA, séptima edición)

Díaz Macías, E., & Guanin Fajardo, J. (2020). Modelo lineal mixto generalizado para la detección de estudiantes exitosos del primer año de Ingeniería de Sistemas. ***Revista Conrado, 16(73)***, 138-142.

RESUMEN

La importancia de anticipar los resultados educativos es esencial para intervenir y controlar los factores que conducen al fracaso académico. Es importante que las instituciones de educación superior dispongan de herramientas que sirvan para el análisis en profundidad de los datos de los estudiantes. Para ello, se evaluó la influencia de variables vinculadas al paso del primer año de la carrera de ingeniería de sistemas. Así, se utilizó el conjunto de datos de estudiantes de 4 períodos de estudio con 5.325 observaciones y 8 variables categóricas. Por lo tanto, se utilizó el modelo mixto lineal generalizado con variable de respuesta binominal (éxito/fallo) con efectos aleatorios. Los resultados de las variables vinculadas al éxito académico se relacionan con el modelo obtenido, a través del contraste de modelos fijos y aleatorios que pueden afectar a los estudiantes exitosos. La validación de los modelos se llevó a cabo mediante la prueba Anova tipo III. A la luz de los resultados, los hallazgos sugieren un modelo significativo, relacionado con los tertiles de las calificaciones del primer parcial, cuartil de asistencia a clases, asignatura y cuerpo docente. Además, proporcionan a los administradores de la educación una visión oportuna de la toma de decisiones.

Palabras clave:

Análisis educativo, educación superior, modelo lineal mixto generalizado, rendimiento académico.

ABSTRACT

The importance of anticipating educational outcomes is essential to intervene and control the factors that lead to academic failure. It is important for higher education institutions to have tools that serve for the in-depth analysis of student data. In order to do so, the influence of variables linked to the passing of the first year of the systems engineering degree was evaluated. Thus, the data set of students from 4 study periods with 5325 observations and 8 categorical variables was used. Hence, the generalized linear mixed model with binominal response variable (success/failure) with random effects was used. The results of the variables linked to academic success are related to the model obtained, through the contrast of fixed and random models that may affect successful students. The validation of the models was carried out by means of the Anova type III test. In the light of the results, the findings suggest a significant model, related to the tertile of grades of the first partial, quartile of attendance to classes, subject and teachers body. In addition, they provide timely insight into decision making to educational administrators.

Keywords:

Educational analysis, higher education, generalized mixed linear model, academic performance.

INTRODUCCIÓN

La Educación Superior como pilar fundamental del desarrollo e innovación científica de los países aporta destacados avances en diversas ramas de la ciencia y tecnología. Las Instituciones de Educación Superior actualmente se encuentran en una escalada de desafíos académicos, sin descuidar el prestigio universitario de acuerdo al ranking mundial de universidades[1]. Los retos educativos pueden ser abordados desde diferentes perspectivas, sin embargo, la diversidad de programas de estudios hace diversificar los resultados en cuanto al estado académico del estudiante. Los sistemas educativos no deben permanecer ajenos a estos retos para enfrentarse al reto de preparar a su alumnado de acuerdo con las exigencias tanto empresarial como industrial (Vilà, Aneas & Rajadell, 2015).

De esta manera, la evaluación del alumnado basado en un proceso de recogida de información útil, relevante y que permite promover mejoras en los procesos de enseñanza-aprendizaje. Ya que la implicación del alumnado en el proceso educativo que tiene lugar en las aulas, con vistas a la mejora del aprendizaje, y alejarse de la intención medidora o sancionadora que a menudo se asocia a los procesos de evaluación, a modo de mera certificación del éxito o fracaso en los procesos (San Martín Gutiérrez, Jiménez Torres, & Jerónimo Sánchez-Beato, 2016). Por otro lado, El fenómeno del abandono de los estudios en el nivel universitario ha visto incrementada su magnitud y relevancia a raíz de la democratización del acceso a la educación superior acaecida a lo largo de la segunda mitad del siglo xx (Esteban García, Bernardo Gutiérrez, & Rodríguez-Muñiz, 2016).

Van der Zanden (2018), en su trabajo indica que los estudiantes de bajo rendimiento y bajo nivel de ajuste experimentaban obstáculos en su vida social, mientras que los estudiantes de rendimiento medio y alto nivel de ajuste experimentaban menor interferencia de sus estudios en su vida social. Los autores de acuerdo con sus resultados indicaron que el éxito de los estudiantes es un concepto multidominio, con subgrupos de estudiantes de primer año que muestran patrones específicos de éxito. Por otro lado, Montmarquette, Mahseredjian & Houle (2001), en su investigación destacan la influencia de variables relativas al rendimiento, tanto el rendimiento académico previo como el rendimiento en la universidad, que correlacionan con la permanencia del alumno en la titulación. En el caso de los abandonos más tempranos, tras el primer semestre de universidad. Los autores también encuentran relación entre el tiempo de dedicación del alumno al trabajo (sea voluntario o remunerado) y el abandono, teniendo mayor probabilidad de abandono aquellos alumnos que mayor tiempo dedican a este tipo de actividades.

Con base en los resultados obtenidos por otros investigadores, este estudio pretende realizar un análisis, de forma comparativa, de diversas variables relacionadas con el estudiantado de la carrera de ingeniería de sistemas y el profesorado. El objetivo de este trabajo es conseguir un modelo predictivo con factores aleatorios o fijos que permita anticipar la superación o fracaso del curso escolar del estudiante de la carrea de ingeniera en sistemas.

MATERIALES Y MÉTODOS

El estudio propuesto dispone del conjunto de datos relacionado con la carrera de ingeniería de sistemas de la universidad en estudio. Proponemos el uso de los modelos mixtos generalizados usando efectos fijos y aleatorios. Varios de los modelos que se comprueban serán sometidos a pruebas de rendimiento de acuerdo con métodos de comprobación.

Conjunto de datos

El conjunto de datos que hemos extraído para el presente estudio está relacionado con los estudiantes del primer año de la carrera de ingeniería en sistemas, cada ciclo de estudio se divide en dos semestres de clases presenciales, siendo cada semestre subdividido en dos parciales. Cada parcial tiene una puntuación de 1.5 y al final el estudiante se presenta al examen final valorado en 4 puntos. El estudiante es evaluado en una escala de 0 hasta 10, siendo 7 la nota aprobatoria.

Tabla 1. Descripción del conjunto de datos de estudiantes del primer año.

Variables	Tipo	Descripción
Periodo	Factor	Curso académico
Mat_nombre	Factor	Nombre de la asignatura
Mat_numero_creditos	Factor	Número de créditos asignados a la materia según el programa de estudio
Semestre_materia	Factor	Primer o segundo semestre
Pro_apellidos	Factor	Descripción del profesorado
Cuartil_notas	Factor	Notas de las asignaturas
Asistencia_total_uno	Factor	Porcentaje de asistencias a clases dividido en 5 categorías
Class	Factor	Fracaso o éxito del curso universitario

La base de datos posee 5325 registros. En primer lugar, hemos realizado un estudio exploratorio de las variables, ya que las 35 variables iniciales del conjunto de datos

[1] http://www.shanghairanking.com/es/arwu2015.html

original no estuvieron aptas para el estudio. Esto, debido a que presentaban muchos datos incompletos. Así, el filtrado de variables relevantes para obtener el modelo predictivo fue reducido en 8 variables que la presentamos en la Tabla 1. En segundo lugar, para evitar el sesgo de las predicciones del modelo, hemos optado por equilibrar las observaciones utilizando un método para el tratamiento de datos desequilibrados (Chawla & Bowyer, 2002)Ripper and a Naive Bayes classifier. The method is evaluated using the area under the Receiver Operating Characteristic curve (AUC. En tercer lugar, se obtienen los modelos mixtos mediante el paquete estadístico *lme4* (Bates, et al., 2015) del programa estadístico R (R Core Team, 2018), posteriormente hemos valorado los resultados.

Se han considerado 12 modelos mixtos con las combinaciones de variables fijas y aleatorias para el análisis, también distintas combinaciones de variables para lograr un modelo que permita predecir el fracaso o éxito del estudiante del primer año de la carrera de ingeniería en sistemas, utilizando el modelo lineal mixto generalizado.

Para la comprobación de los modelos mixtos generalizados se realiza la prueba del test ***anova***. Posteriormente, se comprueba nuevamente mediante el test de Anova pero esta vez con la comprobación del error tipo III.

RESULTADOS Y DISCUSIÓN

El propósito de la investigación, fue emplear el modelo lineal mixto generalizado para la detección de estudiantes exitosos del primer año de ingeniería de sistemas. Como inicio de la investigación hemos realizado el estudio exploratorio de tres variables del conjunto de datos donde se usó en análisis de correspondencia múltiple (Lê, Josse & Husson, 2008).

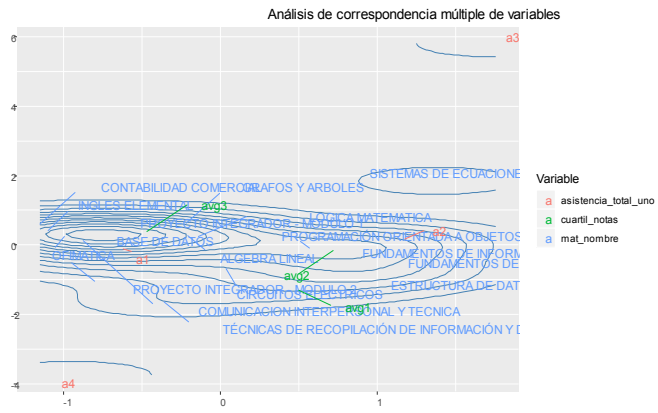

Figura 1. Análisis de correspondencia múltiple de las variables

Se han graficado 3 (Figura 1) variables de interés que se distinguen en grupos que están separados por las líneas de contorno. Se ha detectado concentración de las categorías de cuartiles de asistencia total, es decir, por encima del 50% de asistencia se presentan en contornos diferentes (a3 y a4), sin embargo, es común en los cuartiles de asistencia inferior al 50% (a1 y a2) se ha detectado un conglomerado de categorías de asignaturas y calificaciones.

Modelos obtenidos

De los 12 modelos conseguidos aleatoriamente con distintas combinaciones de factores que ayuden a encontrar el mejor modelo predictivo. Esto, lo hemos realizado mediante la prueba de ***anova***. Finalmente, el ***Modelomixto94*** tuvo mejor valoración en las pruebas.

Hemos conseguido 12 modelos aleatorios con distintas combinaciones de factores (fijos y aleatorios) para encontrar el mejor. De esta manera, la verificación del mejor modelo la hemos realizado mediante la prueba de ***anova***. Finalmente, se ha conseguido que el ***Modelomixto94*** haya tenido mejor valoración en las pruebas del modelo.

Tabla 2. Valoración de modelos mixtos a través de prueba de anova.

	Df	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
Modelomixto93	4	7606.4	7633.9	-3799.2	7598.4				
Modelomixto94	5	7256.9	7291.3	-3623.5	7246.9	351.5031	1	< 2.2e-16	***
Modelomixto90	5	7274	7308.4	-3632	7264	0	0	1	
Modelomixto88	5	7257.9	7292.3	-3624	7247.9	16.0931	0	< 2.2e-16	***
Modelomixto95	6	7162.3	7203.6	-3575.1	7150.3	97.6687	1	< 2.2e-16	***
Modelomixto89	7	7164	7212.1	-3575	7150	0.3047	1	0.581	
Modelomixto98	8	7163.3	7218.4	-3573.7	7147.3	2.6273	1	0.105	
Modelomixto97	8	7161.6	7216.7	-3572.8	7145.6	1.7141	0	< 2.2e-16	***

Modelomixto96	8	7148.1	7203.1	-3566	7132.1	13.5391	0	< 2.2e-16	***
Modelomixto92	9	7217.7	7279.7	-3599.9	7199.7	0	1	1	
Modelomixto91	14	7224.5	7320.8	-3598.2	7196.5	3.2454	5	0.6622	
Modelomixto99	21	7157.9	7302.4	-3557.9	7115.9	80.6054	7	1.04E-14	***

La Tabla 2 muestra las métricas de la prueba de ***anova*** donde el ***Modelomixto94*** es el que mejor valoración tuvo en esta prueba. Aunque la métrica AIC denotó una mejor valoración para el Modelomixto93 este no superó las posteriores pruebas realizadas para estimar su rendimiento. Así mismo con los resaltados en la tabla que tampoco superaron el rendimiento del ***Modelomixto94.***

Los intervalos de confianza (Tabla 3) han mostrado resultados alentadores para este análisis ya que en el intervalo de 2.5% no se muestran valores inferiores a 0, por lo tanto, afianzamos la significación del modelo de predicción para los estudiantes de la carrera de ingeniería en sistemas. Para lograr una mejor interpretación del modelo predicho se muestra la figura 2 que permite tener una mejor interpretación de las predicciones (Breheny & Burchett, 2015).

Tabla 3. Intervalos de confianza del modelo

Variables	2.5 %	97.5 %
pro_apelidos	0.3352132	0.5741053
mat_nombre	0.2834075	0.6514913
asistencia_total_uno	0.2600890	1.5825028
cuartil_notas	0.4773762	2.7247608

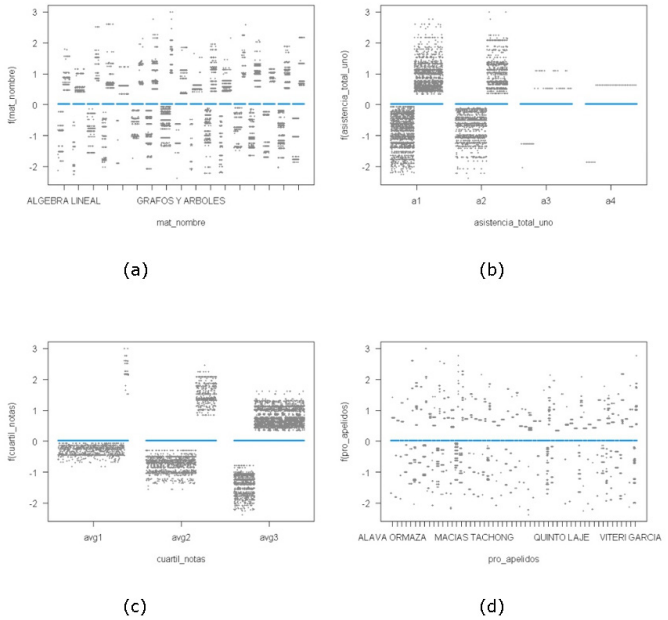

Figura 2. Visualización del modelo predicho ajustado.

La figura (a) relacionada con las asignaturas, (b) asistencia total del primer parcial de clases, (c) calificaciones categorizadas del primer parcial, (d) profesores que impartieron clases en el primer parcial. En todas las figuras se estableció una línea entrecortada que divide la figura en dos zonas, mayores que 0 (positivos-éxito) y menores a 0 (negativo-fracaso).

Los hallazgos encontrados y de acuerdo con el modelo que mejor predijo respecto a los datos hemos considerado que tanto las variables relacionadas con la asistencia de estudiantes, nota, asignaturas y docentes que imparten clases en el primer parcial de la carrera son determinantes para la predicción del fracaso o superación del curso académico. Las métricas usadas para la comprobación del modelo demuestran que la Devianza explicada es (%) = 46,18. Es decir, que las observaciones del conjunto de datos aportan un 46% de información al modelo. Por otro lado, el test de Anova tipo III tiene un p-value de 0,9172.

Los hallazgos nos conducen a tres ámbitos de interés. La primera, relacionada con asignaturas profesionalizantes de la carrera, estas son más difíciles de llevar, aunque las asignaturas de base de datos, fundamentos de informática y estructura de datos, han demostrado ser más llevaderas que el resto. La segunda, relacionada con la asistencia a clases, que por lo general el estudiantado que cumple con un 75% del total de asistencia a clases tiene un grupo mayoritario. La tercera, está relacionado con los promedios, pocos estudiantes aprueban con la nota avg2 (entre 3 y 6), mientras que, un gran grupo lo hace con avg3 (6.1 y 10). Aunque puede no ser concluyente, pero el estudiante mejora la nota cuando se presenta a exámenes extraordinarios de la asignatura (supletorio/recuperación), esto por las normativas propias de la Institución de Educación Superior. Por último, lo relacionado con el profesorado que puede ser un tema de estudio más profundo, sin embargo, lo que destaca es la experiencia del docente en clases.

CONCLUSIONES

El modelo lineal mixto generalizado es un referente imprescindible actualmente en el análisis de datos de investigaciones que pretenden la explicación de fenómenos probabilísticos. Las peculiaridades GLMM descritas son interesantes y permite acoplar las características de las variables al objetivo del estudio, lo cual viene a resolver

el tratamiento estadístico inadecuado en el análisis de datos de investigaciones educativas, donde sucede con frecuencia que las variables que se registran no cumplen los supuestos de los modelos estadísticos tradicionales. Por otro lado, el modelo mixto generalizado, en cualquier caso, es una buena representación de la realidad de los estudiantes de la carrera de ingeniería de sistemas, existen otras técnicas que son necesarias en investigaciones futuras estudiarlas y que están vinculadas con la minería de datos educativos. Independientemente del tipo de enseñanza- aprendizaje, una de las conclusiones más interesantes es la importancia de asistir a clases y el uso del examen supletorio para incrementar el rendimiento académico en los estudiantes de primer curso universitario de la carrera de ingeniería de sistemas.

REFERENCIAS BIBLIOGRÁFICAS

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. ***Journal of Statistical Software***, ***67***(1), 1–48.

Breheny, P., & Burchett, W. (2015). visreg: Visualization of Regression Models. ***The R Journal***, ***9***(2), 56–71.

Chawla, N., & Bowyer, K. (2002). SMOTE: Synthetic Minority Over-sampling Technique Nitesh. ***Journal of Artificial Intelligence Research***, ***16***, 321–357.

Esteban García, M., Bernardo Gutiérrez, A. B., & Rodríguez-Muñiz, L. J. (2016). Persistence in university studies: The importance of a good start. ***Aula Abierta***, 44, 1–6.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. ***J. of Statistical Software***, ***25***(1), 1–18.

Montmarquette, C., Mahseredjian, S., & Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. ***Economics of Education Review***, ***20***(5), 475–484.

R Core Team, D. C. (2018). A Language and Environment for Statistical Computing. ***R Foundation for Statistical Computing***. https://www.gbif.org/es/tool/81287/r-a-language-and-environment-for-statistical-computing.

San Martín Gutiérrez, S., Jiménez Torres, N., & Jerónimo Sánchez-Beato, E. (2016). La evaluación del alumnado universitario en el Espacio Europeo de Educación Superior. ***Aula Abierta***, ***44***(1), 7–14.

Van der Zanden, P. J. A. C., Denessen, E., Cillessen, A. H. N., & Meijer, P. C. (2018). Patterns of success: first-year student success in multiple domains. ***Studies in Higher Education***, 1–15. https://doi.org/10.1080/03075079.2018.1493097

Vilà, R., Aneas, A., & Rajadell, N. (2015). La evaluación de competencias del alumnado en las prácticas externas. la perspectiva de todos los agentes implicados en las prácticas externas del grado de pedagogía de la Universidad de Barcelona. ***Procedia - Social and Behavioral Sciences***, ***196***, 226–232.