Learning Analytics as an analysis factor of university academic performance

Daysi García-Tinisaray[1][0000-0002-7128-5432] José Luis Pino Mejias[2][0000-0001-9344-9242] and Juan Manuel Muñoz Pichardo[3][0000-0001-8841-1987]

[1] Universidad Técnica Particular de Loja, San Cayetano Alto, Ecuador
[2] Universidad de Sevilla, S. Fernando. 4, 4100, España
[3] Universidad de Sevilla, S. Fernando. 4, 4100, España
dkgarcia@utpl.edu.ec

Abstract. The main objective of this research is to use the Learning Analytics approach to identify the covariates that influence the academic performance of university students. There is a multilevel analysis of two levels, in the first level there are 23,583 units of analysis (number of observations-students) and 468 units in the second one (number of groups-classrooms). The results show that the highest percentage of variability is explained by level 1 (students), that all the variables of the Learning Analytics approach have a positive influence and that the participation in chats, forums and video-collaborations cause the greatest impact since they provoke an increase of between 1 and 2 points in academic performance.

Keywords: Learning Analytics, Academic performance, Multilevel.

1 Introducción

El análisis de datos está en auge en el área de la educación sobre todo porque existen herramientas para procesar un volumen creciente de datos, facilitando de esta forma el uso de la información relacionada con el estudiante, el docente, la entidad educativa, etc. con fines de mejorar el aprendizaje.

Las plataformas de enseñanza virtual tales como WEbCT, Moodle, Blackboard, Claroline, Dokeos y recientemente las plataformas MOOC (Massive Open Online Courses) permiten a las universidades monitorizar en tiempo real la actividad de los estudiantes. La integración de esta información con otras variables está en el origen de las técnicas de extracción de conocimiento útil para la mejora del proceso de enseñanza – aprendizaje, conocidas como análisis del aprendizaje (learning analytics).

En materia de rendimiento académico en la educación superior, la mayoría de las investigaciones relevantes presentan un marcado interés en la inclusión de factores personales, son pocos los estudios que hacen un abordaje multinivel que incluya variables del enfoque learning analytic. Los modelos multinivel son más aplicables en el campo educativo porque en estas poblaciones las observaciones individuales no son completamente independientes, es decir se presenta una estructura jerárquica, por lo que según [1], esto implica una dependencia de las observaciones de nivel micro (alumnos) dentro del nivel macro (aulas o centros). Esta dependencia se refiere a que los estudiantes del

mismo grupo comparten el mismo ambiente, mismos profesores, normas, comunicación, etc. A diferencia de la regresión clásica, los modelos multinivel permiten incluir en una misma ecuación, variables independientes de diferentes niveles de agregación.

Bajo estas premisas el objetivo central de esta investigación es emplear el enfoque Learning Analytics para identificar los factores y covariables que influyen en el rendimiento académico de los estudiantes universitarios. Se plantean dos preguntas básicas: ¿Qué proporción de la variación en el rendimiento académico puede atribuirse a las variables que engloba el Learning Analytics? ¿Existe una relación entre el rendimiento académico y el contexto de los estudiantes?

Al identificar la influencia que ejercen sobre el rendimiento académico las variables consideradas (dentro de las cuales consta un grupo de variables de interacción: participación en foro, chat, video-colaboración, número de mensajes enviados al profesor, número de comentarios en el curso de la asignatura, número de accesos al LMS), esta investigación se convierte en un punto de partida de procesos de retroalimentación educativa que permitirán a las instituciones mejorar la focalización de las intervenciones y los servicios de apoyo a estudiantes con mayor riesgo de fracaso académico.

Los resultados del análisis multinivel indican que las variables del nivel 1: edad, rinde supletorio, repite materia, participa en chat, participa en foro, participa en video-colaboración, N° comentarios, N° accesos al LMS y las variables del nivel 2: tasa de repetidores, ciclo y tipo de docente son estadísticamente significativas.

Este artículo está estructurado en seis secciones. La segunda sección contiene la revisión de la literatura. La tercera sección presenta la metodología. En la cuarta sección se presentan los resultados. En la quinta sección se encuentra la discusión de resultados. Finalmente, en la sexta sección constan las conclusiones.

2 Revisión de la literatura

En la actualidad, los enfoques de análisis de datos más usados en el ámbito de la educación superior son la minería de datos educativos (del inglés, Educational Data Mining, EDM), el análisis académico (del inglés, Academic Analytics, AA) y el análisis del aprendizaje (del inglés, Learning Analytics, LA).

El análisis del aprendizaje, análisis académico y minería de datos se centran específicamente en herramientas y métodos para la exploración de datos que provienen de contextos educativos [2]. Hoy en día se considera que estas técnicas ayudan a moldear el futuro de la educación superior y a generar nuevos enfoques y estrategias en mejora de la enseñanza y del aprendizaje.

La diferencia entre estos tres enfoques se establece en los siguientes planteamientos [3]:

- La minería de datos es un desafío técnico ¿Cómo se puede extraer valor de los grandes conjuntos de datos relacionados con el aprendizaje?
- El análisis del aprendizaje es un desafío educativo ¿Cómo se puede optimizar las oportunidades para el aprendizaje en línea?

- El análisis académico es un desafío económico / político ¿Cómo se puede mejorar sustancialmente las oportunidades de aprendizaje y los resultados educativos a nivel nacional o internacional?

Estos enfoques no solo recogen y exploran grandes cantidades de información, sino que permiten construir y poner a prueba modelos que se centran en el estudiante, ya sea de forma individual o en el contexto de la institución, con la finalidad de predecir o mejorar el rendimiento académico.

El Learning Analytics surge a partir de dos tendencias convergentes: el uso cada vez mayor de los Entornos Virtuales de Aprendizaje en las instituciones educativas y la aplicación de técnicas de minería de datos para los procesos de inteligencia de negocios en sistemas de información de la organización [4].

"Learning Analytics es la medición, recopilación, análisis y presentación de datos sobre los alumnos y sus contextos, a efectos de entender y optimizar el aprendizaje y los entornos en los que ocurren los sucesos de aprendizaje" [5]

El informe Horizont [6] menciona que el Learning Analytics tiene su origen en la minería de datos aplicada al sector comercial en donde se realizaban análisis de las actividades de los consumidores con la finalidad de personalizar la publicidad.

Este tipo de análisis permite usar los datos asociados con el aprendizaje de los estudiantes y generar informes que sean útiles para los docentes (actividades y progreso de los estudiantes), para los estudiantes (retroalimentación) y para los administradores (incremento de aulas de clase, tasa de graduación, etc.) [7].

3 Metodología

3.1 Fuente de datos

Los datos utilizados provienen de una de las universidades ecuatorianas con más número de estudiantes a nivel de educación superior a distancia en Latinoamérica, a partir de esta información se desarrollan los dos análisis antes mencionados cuya variable objetivo es el rendimiento académico.

La población objeto de estudio comprende un ámbito individual, grupal y contextual, los participantes que la conforman son 23,583 estudiantes y 468 aulas. Los datos se ordenaron jerárquicamente, de tal forma que las observaciones se agrupen correctamente en cada uno de los niveles de agregación. Los datos fueron levantados en el año 2014.

En el proceso de inclusión de variables usadas para la modelización del rendimiento académico se tuvo en cuenta el enfoque de enseñanza centrada en la teoría del Learning Analytics, por lo que se trabaja con datos suministrados por el Entorno Virtual de Aprendizaje, una de las herramientas de apoyo principales en esta modalidad de estudio.

3.2 Variables

Las variables se han seleccionado en pro del cumplimiento de los objetivos específicos. Estas variables son de carácter académico, demográfico, pedagógico y tecnológico (en el ámbito tecnológico se trabaja con variables que involucra el enfoque "learning analytics").

Se toman en cuenta variables individuales del estudiante (nivel inferior), variables del docente y asignatura (nivel intermedio) y variables de la escuela (nivel superior). Todas las variables se obtienen dentro de la misma universidad, de esta forma, se supone que la correlación promedio (conocida como la correlación intraclase) entre las variables de los alumnos de la misma universidad y del mismo tipo de asignatura (troncal) es mayor que la correlación de las mismas variables medidas entre los alumnos de universidades distintas. Estas covariables se presentan en la Tabla 1.

La variable de respuesta se denomina rendimiento académico y es la calificación final del estudiante que se mide en un rango de 0 a 40 puntos (incluye la sumatoria de los exámenes, trabajos a distancia y otras actividades).

Table 1. Variables de estudio

Niveles	Covariables	Dimensión
Nivel 1 Estudiantes (23,583)	Edad	Sociodemográficas
	Género	
	Región	
	Repite la asignatura	Antecedentes académicos
	Rinde supletorio	
	Tiene Beca	
	N° de consultas al profesor	Learning analytics
	N° de comentarios	
	N° de accesos a LMS*	
	N° de accesos asignatura	
	Tiempo de uso LMS	
	Participación en foros	
	Participación en video colaboración	
	Participación en chat	
Nivel 2 Aulas (468)	Número de matriculados	Asignatura
	Número de repetidores	
	Número de créditos	
	Ciclo de asignatura	
	Años de experiencia	Docente
	Evaluación docente	
	Formación académica	
	Tipo de docente	

*Learning Management System.

3.3 Ecuaciones

El inicio de la aplicación de los modelos multinivel en el campo educativo se debe principalmente al aporte que realizaron [8], en su investigación "Statistical modelling issues in school effectiveness studies" en la cual introdujeron por primera vez el análisis multinivel para determinar la efectividad escolar, demostrando la existencia de errores metodológicos al usar las regresiones tradicionales en investigaciones anteriores y reconociendo la presencia de una estructura jerárquica en la presentación y análisis de datos entre estudiantes y escuelas.

Los modelos multinivel han estado aplicándose con mayor fuerza en el campo de la salud y educación desde hace más de dos décadas [9], [10], [11].

El análisis multinivel se desarrolla de acuerdo a la estructura anidada que presente la población en estudio, ésta básicamente suele ser de 2 o 3 niveles. Conforme se aumentan los niveles se supone que aumenta la heterogeneidad.

El modelo de 2 niveles se compone de dos estimaciones en donde i =1,…,nj unidades del nivel 1 se encuentran anidados dentro de j=1,…,J unidades del nivel 2.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + ... + \beta_{nj}x_{nij} + e_{ij} \quad (1)$$

La ecuación (1) representa la modelización del nivel 1, en donde y_{ij} es la variable dependiente para el caso i en la unidad j, β_{nj} es el coeficiente del nivel 1, x_{nij} es la variable explicativa n para el caso i en la unidad j y el efecto aleatorio del nivel 1 se representa por e_{ij}.

$$\beta_{nj} = \gamma_{n0} + \gamma_{n1}W_{1j} + ... + \gamma_{np}W_{nj} + u_{nj} \quad (2)$$

La modelización de 2 niveles se establece en la ecuación (2), en donde β_{nj} es la variable dependiente (coeficientes del nivel 1), γ_{np} coeficientes del novel 2, W_{nj} representa a las variables explicativas p para la unidad j del nivel 2 y u_{nj} es el efecto aleatorio del nivel 2.

4 Resultados

Se realiza un procedimiento "stepwise" hacia adelante, es decir incrementando el número de variables explicativas del nivel 1 y del nivel 2 para ir ampliando la capacidad de explicación y ajuste del modelo, aunque para ello esta secuencia metodológica aumente simultáneamente la complejidad del mismo.

Las estimaciones "stepwise" se desarrollan bajo una especificación lineal debido a las características de los datos y con el apoyo del software Stata/SE 12.0 a través de la funcionalidad Statistic – Multilevel mixed-effects models.

El análisis se inicia con el paso 0 (Modelo nulo-ANOVA con efectos aleatorios) en el cual no se incluyen variables explicativas, es decir, se estima un modelo nulo para comprobar la significatividad y luego explicar la varianza, expandiendo el modelo a través de la incorporación de predictores de los dos niveles en la parte fija y aleatoria

Table 2. Modelo Final

Mixed-effects ML regression			Number of obs =			23,583
Deviance = 161611.82			Wald chi2(16) =			6146.69
Log likelihood = -80805.908			Prob > chi2 =			0.0000
REND_ACADEMICO	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Tasa_Repetidores	-21.2218	1.609523	-13.19	0.000	-24.3764	-18.06719
Ciclo	.7623473	.0697031	10.94	0.000	.6257318	.8989628
Tipo_docente						
Tiempo completo	0	(base)				
Administrativo	1.510935	.5678153	2.66	0.008	.3980378	2.623833
Invitado	.9573145	.3116882	3.07	0.002	.3464169	1.568212
Edad	.0837473	.0062392	13.42	0.000	.0715186	.095976
Rinde_supletorio	-.5267685	.2636077	-2.00	0.046	-1.04343	-.0101069
Rinde_supletorio*Ciclo	-.3366908	.0581579	-5.79	0.000	-.4506782	-.2227034
Repite_materia	2.808255	.2476235	11.34	0.000	2.322922	3.293588
Repite_materia*Ciclo	-.2486661	.0633913	-3.92	0.000	-.3729107	-.1244215
Participa_chat	1.313279	.1692679	7.76	0.000	.98152	1.645038
Participa_foro	2.057453	.1272299	16.17	0.000	1.808087	2.306819
Participa_video	1.31303	.1937892	6.78	0.000	.9332105	1.69285
N_comentarios	.091933	.0222558	4.13	0.000	.0483125	.1355535
N_accesos_LMS	.0438462	.0025096	17.47	0.000	.0389275	.048765
N_acce-sos_LMS*Tasa_Repetidores	.0696397	.0086301	8.07	0.000	.052725	.0865544
N_accesos_LMS*Ciclo	-.0042219	.0003348	-12.61	0.000	-.004878	-.0035658
_cons	20.25.424	.5565587	36.39	0.000	19.1634	21.34.507

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
AULA: Independent				
var(Rinde_supletorio)	3.243811	.4901397	2.412.343	4.361863
var(Repite_materia)	.8369861	.4001639	.3279125	2.13638
var(Partica_chat)	.4351275	.3708907	.0818607	2.312903
var(Participa_foro)	.3788677	.2529417	.1023773	1.402075
var(Participa_video)	.5905212	.4444584	.135075	2.581642
var(_cons)	7.385529	.6798355	6.166359	8.845746
var(Residual)	52.75249	.5014936	51.77868	53.74461

LR test vs. linear regression: chi2(6) = 2003.78 Prob > chi2 = 0.00

Se continua con el paso 1 (Explicación del intercepto con variables del nivel 2) se consideran únicamente predictores del nivel 2, con la finalidad de explicar la variabilidad a través de variables del nivel 2.

Para el paso 2 (Significación de las variables explicativas del nivel 1) se ingresan predictores del nivel 1 y estos son los que explican la varianza del rendimiento académico dentro de los grupos. En el paso 3 (Regresión con interacciones y variables de los niveles 1 y 2) se consideran los resultados anteriores para generar una estimación basada en las variables explicativas de los estudiantes y de las aulas que son estadísticamente significativas y se realizan las interacciones multinivel a nivel del alumno con variables explicativas de las aulas.

Finalmente en el paso 4 (Variabilidad en los coeficientes de los predictores del nivel 1), a diferencia del paso 3, se incluye en la parte de efectos aleatorios las pendientes significativas del nivel 1.

Analizando todas las estimaciones para dos niveles (estudiantes y aulas), las estimaciones que explican un mayor porcentaje de la varianza son las del paso 3 y 4, sin embargo, la estimación que mejor se ajusta es la del paso 4, por lo que es este modelo el que se considera como modelo final definitivo para dar respuesta al objetivo prefijado.

El modelo que resulta en el paso 4 se presenta en la Tabla 2 estos datos muestran que después de incluir las interacciones, el componente de la varianza de las pendientes de las variables explicativas del nivel 1 muestra una variación leve pero significativa entre aulas.

5 Discusión de resultados

El modelo final involucra: tres covariables del Nivel 2: tasa de repetidores, ciclo y tipo de docente. Ocho variables del Nivel 1: edad, rinde supletorio, repite materia, participa en chat, participa en foro, participa en video colaboración, N° comentarios, N° accesos al LMS. Cuatro interacciones multinivel. La varianza de cinco pendientes del Nivel 1.

El coeficiente de la variable tasa de repetidores medida en el intervalo [0,1], nos indica que un aumento en 10 puntos porcentuales de estudiantes matriculados por segunda o tercera vez en una asignatura troncal, ocasiona una disminución de 2.1 puntos en el rendimiento académico. Esto significa que a pesar de que se asume que los estudiantes tienen más experiencia que los estudiantes nuevos en la materia, no obtienen una mejor nota, lo cual podría estar ligado a la metodología de enseñanza o a los instrumentos de evaluación.

Otra variable del nivel 2 es la variable ciclo. Los resultados indican que cuando la asignatura se encuentra en un ciclo superior el rendimiento académico incrementa en 0.8 décimas. Esto se puede esperar ya que se considera que conforme un estudiante avanza a ciclos superiores tiene más conocimientos y en cierta forma ha adquirido madurez académica.

La pendiente de la variable tipo de docente influye positivamente sobre el rendimiento académico, ya que, este tiende a subir aproximadamente 1 punto si el docente es administrativo o invitado. Estos resultados se pueden explicar posiblemente por dos razones: la primera sería que los docentes a tiempo completo son más estrictos y la segunda puede ser que estos docentes tienen más créditos o asignaturas a su cargo en comparación a los docentes invitados o administrativos. Estos en sí son dos supuestos, que se deberían de verificar en base a otros aspectos como salario que perciben, número de asignaturas que tienen a su cargo, años de experiencia, etc.

En cuanto a la edad, los resultados indican que por un año más de edad que tenga el estudiante, el puntaje del rendimiento académico subirá en 0.08 décimas. El comportamiento de estos resultados coinciden con los planteados por [12], [13], quiénes encontraron que la edad tiene una relación positiva y significativa con el rendimiento académico de los estudiantes universitarios.

El coeficiente de la pendiente de la variable rinde supletorio y su interacción con el ciclo indican que si un estudiante se queda suspenso y está en un ciclo superior el rendimiento académico disminuirá en 0.86 décimas (resultante de la suma de los coeficientes -0.52677 y -0.33669 recogidos en la Tabla 2). Mientras que analizando los resultados de la variable repite materia y su interacción con el ciclo nos muestra que si un estudiante repite la materia y está en un ciclo superior, el rendimiento académico en promedio subirá en 2.6 décimas (resultante de la suma de los coeficientes 2.80826 y -0.24867 recogidos en la Tabla 2).

Todas las variables del enfoque Learning Analytics tienen una relación positiva con el rendimiento académico, siendo la participación en chat, foro y video-colaboración las que ocasionan el mayor impacto ya que provocan un incremento de entre 1 y 2 puntos en el rendimiento académico, afirmando de esta forma que si existe una relación significativa con el rendimiento académico tal como lo plantean [4], [14]. La variable N° accesos al LMS interacciona con la tasa de repetidores y el ciclo de la asignatura, lo cual indica que ocasiona un incremento de cerca de 0.11 décimas en el rendimiento académico (resultado de la suma de los coeficientes 0.04385, 0.06964 y -0.00422 recogidos en la Tabla 2).

6 Conclusiones

Las variables incluidas en la presente investigación permiten identificar cual es la influencia que ejercen sobre el rendimiento académico, estas estimaciones pueden permitir a una institución educativa mejorar la focalización de las intervenciones y los servicios de apoyo a estudiantes en riesgo de problemas académicos.

Los resultados obtenidos dan respuesta a las hipótesis y objetivos planteados, además este trabajo es un punto de partida para futuras investigaciones que consideren que el ámbito tecnológico se está convirtiendo en una de las mejores herramientas de enseñanza aprendizaje, sobre todo en educación a distancia

Todas las variables del enfoque Learning Analytics tienen una influencia positiva sobre el rendimiento académico de estudiantes universitarios, específicamente la participación en chats, foros y video-colaboraciones ocasionan el mayor impacto ya que provocan un incremento de entre 1 y 2 puntos en el rendimiento académico.

References

1. Snijders, T., Bosker, R.: Standard errors and sample sizes for two-level research. Journal of educational statistics, 18(3), 237–259 (1993).
2. Dyckhoff, A., Zielke, D., Bültmann, M., Chatti, M., Schroeder, U.: Design and Implementation of a Learning Analytics Toolkit for Teachers. Journal of Educational Technology & Society, 58–76 (2012).
3. Ferguson, R.: Learning analytics: drivers, developments and challenges International Journal of Technology Enhanced Learning, 304–317 (2012).
4. Agudo, A., Hernandez, A., Iglesias, S.: Predicting academic performance with learning analytics in virtual learning environments: a comparative study of three interaction classifications. 2012 International Symposium on Computers in Education (SIIE), pp. 1–6. IEEE Xplore, Andorra la Vella (2012).
5. Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Buckingham, S., Ferguson, R.: Open Learning Analytics: an integrated & modularized platform Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. Obtenido de http://solaresearch.org/OpenLearningAnalytics.pdf
6. Johnson, L., Smith, R., Willis, H., Levine, A., Haywood, K. The 2011 Horizon Report. Homepage, de http://net.educause.edu/ir/library/pdf/hr2011.pdf, last accessed 2015/10/15.
7. Brown, M.: Learning Analytics: the coming third wave. EDUCAUSE Learning Initiative Brief, 1 (4), 1–4 (2011).
8. Aitkin, M., Longford, N.: Statistical modelling issues in school effectiveness studies. Journal of the Royal Statistical Society, 1–43 (1986).
9. Goldstein, H., Spiegelhalter, and D.: League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society, 385–443 (1996).
10. Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D.: A multilevel analysis of school examination results. Oxford review of education, 425–433 (1993).
11. Draper, D.: Inference and hierarchical modeling in the social sciences (with discusion). Journal of Educational and Behavioral Statistics, 115–147(1995).
12. Nasir, M.: Demographic characteristics as correlates of academic achievement of university students. Academic Research International, 400–405 (2012).
13. Alhajraf, N., Alasfour, A.: The impact of demographic and academic characteristics on academic performance. International Business Research, 92–100 (2014).
14. Yu, T., Jo, I.: Educational Technology Approach toward Learning Analytics: Relationship between Student Online Behavior and Learning Performance in Higher Education. ACM International Conference Proceeding Series, 269–270 (2014).