Fernando Martínez Abad fma@usal.es *Salamanca, 20 de abril de 2017*
Instituto Universitario de Ciencias de la Educación, Universidad de Salamanca

Detección de buenas prácticas educativas en escuelas de alto valor añadido mediante técnicas de *Big Data*

Descripción del proyecto

El objetivo principal del presente proyecto es la **detección de factores asociados al rendimiento en escuelas de alto valor añadido para la elaboración y difusión a la comunidad de un catálogo de buenas prácticas educativas**. De este modo, el proyecto tendrá una extensión de 15 meses, a lo largo de los que se abordarán cronológicamente 3 objetivos clave:

1. **Aplicación de modelos jerárquicos lineales para la detección de escuelas de alto y bajo valor añadido en base a evaluaciones a gran escala**

Dado que la realidad del sistema educativo en general, y de las escuelas en particular, es compleja, variada y asociada a múltiples factores, se requiere para su análisis un enfoque amplio que tenga en cuenta todo el entorno y que sea capaz de considerar las diferencias contextuales y estructurales existentes entre la multitud de ámbitos en los que se desarrolla. Es por esto que se ha generalizado en los últimos años un enfoque mixto en el estudio del rendimiento académico y sus factores asociados (Chapman, Muijs, Raynolds, Sammons, & Teddlie, 2016). Con este fin, se han venido perfeccionando una serie de técnicas estadísticas de naturaleza multivariante, denominadas modelos jerárquicos lineales o modelos multinivel, que hacen una estimación más ajustada de la aportación específica de los centros educativos a la formación de los estudiantes (Bryk & Raudenbush, 1992; Goldstein, 1995), lo cual facilita la detección de escuelas con una eficacia sensiblemente superior (o inferior) a la previsible dadas las condiciones contextuales de partida propias y de sus estudiantes. Gracias a estas técnicas, es posible eliminar la influencia que factores contextuales significativos (covariables) tienen sobre el rendimiento de estudiantes y escuelas, obteniendo un índice que informa sobre la parte del rendimiento promedio de las escuelas no explicada por estos factores contextuales. Y es este rendimiento promedio residual obtenido en los modelos jerárquicos lineales el que se viene denominando, de manera generalizada, como Valor Añadido en Educación (Martínez Abad, Lizasoain Hernández, Castro Morera & Joaristi Olariaga, 2017).

En esta primera etapa del proyecto se procederá a la aplicación de modelos multinivel a partir de la muestra de estudiantes (n=37.205) y centros educativos (m=980) españoles del programa de evaluación a gran escala PISA[1] 2015, incluyendo como variable dependiente el rendimiento académico de los estudiantes y como variables independientes las variables contextuales significativas[2]. Así, dado que en las pruebas PISA se evalúa el rendimiento en tres áreas específicas por separado (Lectura, Matemáticas y Ciencias), se generarán un total de 3 modelos a partir de la base de datos seleccionada, por lo que se obtendrán, por centro educativo, 3 medidas del rendimiento promedio residual (aporte de la escuela al rendimiento del estudiante tras eliminar los efectos de contexto). Tras la obtención de estas medidas residuales del aporte o valor añadido de las escuelas, se seleccionarán las escuelas de un mayor y menor valor añadido que cumplan los siguientes criterios:

[1] www.oecd.org/**pisa**/

[2] Las principales variables contextuales incluidas en las evaluaciones de contexto de las pruebas PISA, y que resultan asociadas al rendimiento académico de manera generalizada en el estado de la cuestión son: Índice Socio-Económico, Recursos en el hogar, Recursos en la escuela, Condición de repetidor, Condición de inmigrante, Género, Tamaño del aula, Mes de nacimiento y Haber cursado la educación infantil.

- Superar un percentil mínimo/máximo en el residuo promedio de los 3 modelos
- Alcanzar un percentil mínimo/máximo en los residuos de cada uno de los 3 modelos

Dado que se planifica seleccionar, de entre los 980 centros de la muestra, un 5% de centros con valor añadido (2.5% de alto valor añadido y 2.5% de bajo valor añadido), y que únicamente se seleccionarán centros con una muestra de estudiantes en PISA 2015 superior o igual a 15, los percentiles exactos utilizados en esta fase se determinarán durante el mismo proceso de selección. Tras este proceso está previsto, por tanto, seleccionar un total de 24 centros educativos de bajo valor añadido (de eficacia claramente inferior a la previsible dadas sus características contextuales) y 24 centros de alto valor añadido (de eficacia claramente superior a la previsible dadas sus características contextuales).

Esta fase finaliza con una caracterización inicial de los centros seleccionados como medida de control y evaluación del proceso, estudiando si se ha obtenido un listado de centros variado, tanto en su ubicación (comunidad autónoma), como en su titularidad, tamaño o localización rural/urbana. El no cumplimiento de esta condición llevará a un reajuste del proceso anterior para asegurar la variedad y multiplicidad necesaria en todo estudio propio del ámbito educativo.

2. Aplicación de técnicas de *Big Data* para la detección de factores asociados al rendimiento en las escuelas de alto valor añadido

El rendimiento académico es un fenómeno multicausal en el que intervienen infinidad de variables íntimamente relacionadas entre sí. Identificar estas variables es una labor compleja, debido tanto a su diversidad como a la interrelación que existe entre ellas (Lee & Shute, 2010). Esta realidad, unida a la necesidad de discriminar entre información valiosa y accesoria en un contexto de datos educativos masivos, lleva a la propuesta de esta segunda fase del proyecto, en la que se implementarán técnicas multivariantes de *Big Data* a partir de las bases de datos de las evaluaciones a gran escala PISA 2015, para identificar las variables predictoras no contextuales determinantes en los centros educativos identificados en la anterior fase como de alto valor añadido. Dicho de manera más simple, en esta fase de la investigación se aplicarán algunas técnicas propias del *Big Data* para identificar, de entre el conjunto de variables no contextuales incluidas en las pruebas PISA 2015[3], aquellos factores que sean capaces de discriminar de manera más significativa las escuelas de alto valor añadido de las escuelas de bajo valor añadido.

De este modo, se pretende alcanzar un alto grado de innovación al integrar técnicas que actualmente se están aplicando de manera generalizada en el ámbito educativo a partir de estudios a gran escala (modelos jerárquicos lineales), junto con otras técnicas que, en una segunda fase, permitan extraer información valiosa a partir de los centros identificados en la primera fase. Esta información alcanza un alto valor para la comunidad educativa y para las instituciones públicas, ya que aporta evidencias en relación a los factores que contribuyen de manera determinante en la mejora del rendimiento académico de los estudiantes. De este modo, los resultados obtenidos en este proyecto harán un aporte importante no sólo en los micro y meso-contextos educativos (familias, profesorado, equipos

[3] Cabe recordar que las bases de datos de las pruebas PISA 2015 incluyen cientos de factores no contextuales, asociados a las escuelas (autonomía escolar, liderazgo, clima escolar, actividades extraescolares, proyectos de centro, etc.), asociados a los profesores (participación del profesorado en la escuela, desarrollo profesional, colaboración, satisfacción, etc.) y asociados a los estudiantes (autoeficacia, motivación, actitudes, trabajo en equipo, horas de estudio, etc.).

directivos y escuelas), indicando a los distintos agentes educativos algunas cuestiones esenciales que pueden promover una mejora de la eficacia escolar, sino que pueden proveer un buen punto de partida para orientar políticas educativas institucionales a nivel macro-contextual, que busquen fomentar algunos factores clave aquí identificados.

Cabe destacar que, a pesar de la versatilidad que ofrecen las técnicas relacionadas con el *Big Data*, también conocidas en el ambiente científico como *Minería de Datos* (Data Mining), estos procedimientos se utilizan de forma escasa en el estudio de factores asociados al logro académico, sobre todo en niveles de educación obligatoria, donde su aplicación es casi nula (Kiray, Gok & Bozkir, 2015; Hsieh, 2013, Martínez Abad & Chaparro Caso López, 2017). A pesar de que existen pocas evidencias en el estado de la cuestión que nos ayuden en la toma de decisiones posterior sobre el proceso de análisis de datos, nos apoyaremos en la propia experiencia en este tipo de procesos (Martínez Abad & Chaparro Caso López, 2017) para la selección de las técnicas más apropiadas.

En base al estudio del conjunto de técnicas disponibles, y a las necesidades de esta fase de la investigación, se prevé la aplicación de algoritmos de clasificación, en concreto el algoritmo C4.5 (Árboles de Decisión), como extensión del algoritmo ID3. C4.5 genera un árbol ramificado con las variables predictoras a partir de n iteraciones, seleccionando la/s variable/s predictora/s más discriminante/s con respecto al criterio en cada iteración, hasta alcanzar la condición de parada.

En total, está previsto que la base de datos empleada en esta fase integre 48 centros educativos y aproximadamente 2000 estudiantes (conforme a los datos disponibles en las pruebas PISA 2015, se han estimado en promedio unos 40 estudiantes muestreados por escuela). La variable criterio (dependiente) en esta fase es la identificación de la escuela como de alto o bajo valor añadido (variable dicotómica), y se prevé que tras el filtrado de las bases de datos, se disponga de alrededor de 150-200 variables predictoras en la aplicación de los algoritmos. Este proceso dará lugar a la identificación de los 15-20 factores más importantes y al análisis de su diferente comportamiento en las escuelas de alto y bajo valor añadido. Al igual que en la fase anterior, se implementará un procedimiento para controlar y asegurar la consecución del objetivo propuesto en esta fase, que consistirá en el estudio de la bondad de ajuste de los modelos (Kappa de Cohen, error absoluto medio, área ROC y precisión) y de su capacidad de generalización (independencia, cross-validation). En caso de detectarse problemas en el ajuste y la generalización de los modelos, se estudiará la aplicación de otros algoritmos o técnicas diferentes a lo propuesto.

Finalmente, como resultado principal de esta fase, se aprovechará la información obtenida en todo el proceso aplicado para llevar a cabo una caracterización y diferenciación de las escuelas de alto y bajo valor añadido con respecto a sus niveles en las variables predictoras identificadas.

3. Diseño de un catálogo de buenas prácticas educativas y difusión de resultados a la comunidad educativa y científica

Una cuestión esencial en todo proceso de investigación aplicada es la transferencia del conocimiento generado, no sólo a la comunidad científica, sino principalmente a la Sociedad y a la Comunidad implicada, de cara a que se promuevan mejoras directas e indirectas en su entorno. El presente proyecto plantea en esta tercera y última fase el contacto y difusión de los resultados a los diferentes agentes e instituciones educativas a los que puedan resultar de interés. Así, se plantean dos grandes estrategias en esta fase: divulgación a la comunidad educativa y difusión a la comunidad científica.

En cuanto a la divulgación de los resultados a la comunidad educativa, se prevén varias acciones:

- Redacción y difusión, en forma de catálogo, de un manual de buenas prácticas educativas (modelo divulgativo extendido): Destinado principalmente a profesores, equipos directivos e instituciones educativas (nivel meso y macro), este catálogo incluirá información divulgativa, pero exhaustiva, sobre los factores clave de eficacia escolar localizados en el proyecto. El objetivo de este catálogo es alentar y facilitar la puesta en marcha de acciones estratégicas en torno a la mejora de la eficacia escolar por parte de escuelas e instituciones educativas.
- Elaboración y difusión de un catálogo de buenas prácticas resumido, en forma de folleto informativo (modelo divulgativo reducido): Destinado principalmente a familias y otros agentes educativos del ámbito no profesional (nivel micro). Este catálogo incluirá, desde una perspectiva práctica, algunas claves y sugerencias básicas que pueden ayudar a familias y a los propios escolares en la mejora de su rendimiento académico.

Para la difusión de estos catálogos está previsto el contacto directo, por vía postal y/o digital, con distintas asociaciones, instituciones y grupos (centros educativos, AMPAS, Centros de Formación del Profesorado, Administraciones públicas, etc.), a los que se facilitará la información y se invitará, en su caso, a difundirla si lo consideran oportuno.

En lo que respecta a la difusión a la comunidad científica, como evidencias asociadas a los resultados, se establece el compromiso de desarrollar en el marco del presente proyecto la siguiente producción: 3 ponencias en congresos internacionales del ámbito educativo y de eficacia escolar, preferentemente que publiquen actas con ISBN, 1 publicación en revista científica nacional de impacto medio (revista indexada al menos en SCOPUS) y 1 publicación en revista científica internacional de alto impacto (indexada en JCR, en los primeros cuartiles).

En paralelo a este plan de difusión, se elaborará un espacio web del proyecto en el que se irá incorporando información actualizada sobre los resultados alcanzados. Esta Web tendrá una doble vertiente de difusión científica y de divulgación a la comunidad educativa, reuniendo tanto datos sobre las cuestiones más técnicas del proceso y las publicaciones y participación en eventos científicos, como información práctica sobre los factores asociados al rendimiento detectados y cómo aprovecharlos en la práctica educativa concreta. Al respecto, se prevé integrar en la Web un canal de comunicación e interacción a partir del que los distintos agentes y profesionales educativos puedan intercambiar opiniones, ideas y ejemplos de experiencias educativas concretas en el ámbito de la eficacia escolar.

Finalmente, en función de las posibilidades económicas del proyecto, está previsto difundir tanto los catálogos como la información general de la Web tanto en los 4 idiomas cooficiales del país (español, euskera, catalán y gallego) como en lengua inglesa.

Referencias bibliográficas

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park: Sage Publications.

Chapman, C., Muijs, D., Raynolds, D., Sammons, P., & Teddlie, C. (Eds.). (2016). *The Routledge International Handbook of Educational Effectiveness and Improvement Research, policy, and practice*. New York: Routledge.

Goldstein, H. (1995). *Multilevel statistical models*. London; New York: Oxford University Press.

Hsieh, M. (2013). Data mining from education databases examine the factors impacting the school performance in the United States. *International Journal of Intelligent Technologies and Applied Statistics, 6*, 135-143. doi: 10.6148/IJITAS.2013.0602.03

Kiray, S. A., Gok, B., & Bozkir, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health (JESEH), 1*, 28-48.

Martínez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39–55. https://doi.org/10.1080/09243453.2016.1235591

Martínez Abad, F., Lizasoain Hernández, L., Castro Morera, M., & Joaristi Olariaga, L. M. (2017). Selección de escuelas de alta y baja eficacia en Baja California (México). *Revista Electrónica de Investigación Educativa*, *19*(2), 38–53.

Plan de trabajo

Actividades o tareas :	**2017**		**2018**		
	Jul.-Sept.	**Oct.-Dic.**	**Ene.-Mar.**	**Abr.-Jun.**	**Jul.-Sept**
1. Aplicación de Modelos Jerárquicos Lineales para la detección de escuelas de alto y bajo valor añadido					
i. Acceso, adaptación y filtrado de las Bases de Datos PISA 2015					
ii. Testeo, aplicación y obtención de los modelos jerárquicos lineales					
iii. Selección de escuelas de alto y bajo valor añadido					
iv. Caracterización de las escuelas seleccionadas y evaluación de los resultados de la selección					
2. Aplicación de técnicas de *Big Data* para la detección de factores asociados al rendimiento					
i. Preparación de las Bases de Datos					
ii. Aplicación del algoritmo de clasificación y obtención de los árboles de decisión					
iii. Análisis y validación de los árboles de decisión: Identificación de los factores clave asociados al rendimiento					
iv. Descripción de factores existentes y buenas prácticas implementadas en las escuelas de alto valor añadido					
3. Diseño de un catálogo de buenas prácticas educativas y difusión de resultados					
i. Diseño, redacción y traducción de catálogos extendido y reducido de buenas prácticas					
ii. Diseño, desarrollo y traducción de espacio web del proyecto					
iii. Difusión de resultados a la comunidad científica					
iv. Divulgación de buenas prácticas a la comunidad educativa					