# CAPÍTULO 5.
# BIG DATA Y MINERÍA DE DATOS APLICADOS A LA EVALUACIÓN DEL IMPACTO EDUCATIVO

JUAN J. LEIVA OLIVENCIA, PABLO D. FRANCO CABALLERO Y ANTONIO MATAS TERRÓN

*Universidad de Málaga*

## 1. INTRODUCCIÓN

En un mundo interconectado donde la información se ha convertido en la materia prima de la comunicación y también de la educación, la expansión de las redes sociales, así como de los grandes canales, autopistas y "almacenes" digitales está cambiando la fisonomía de las propias producciones y significados humanos en la era digital. Un tiempo impregnado de enorme complejidad e incluso perplejidad por la velocidad e inmediatez de los cambios que acontecen de forma acelerada donde se generan, crean, construyen y reconstruyen datos, informaciones y conocimientos (García-Aretio, 2017).

Todas las personas son creadoras de datos susceptibles de ser compartidos y las propias redes sociales funcionan como telarañas de información diversa que contribuye a complejizar aún más la sociedad del conocimiento en la que vivimos y habitamos como seres históricos, sociales y emocionales en múltiples dimensiones y entornos. Dicho esto, no se puede negar que el entorno o esfera virtual se ha constituido en un eje fundamental que está transformando las relaciones interpersonales, la construcción del conocimiento científico, el desarrollo tecnológico y el propio devenir de las identidades personales y sociales en un mundo completamente digitalizado.

La digitalización de todos los procesos y fenómenos humanos es una creciente oportunidad para el aprendizaje y la mejora del bienestar y calidad de vida de los seres humanos. Obviamente esta afirmación se sitúa en un enfoque social y pedagógico de inclusión, esto es, concibiendo que las innovaciones tecnológicas y el desarrollo digital debe estar al servicio de la humanidad y no viceversa. Es de-

cir, que la acumulación de datos, informaciones y de mejoras que proporciona las redes telemáticas y las plataformas digitales deben favorecer el bienestar y el desarrollo inclusivo de todas las personas y grupos sociales (Serrano-Cobos, 2014).

En el ámbito formativo, se puede identificar un buen número de oportunidades y características propias del Big Data, pero resulta necesario ir clarificando conceptos e ideas para arrojar luz en la posibilidad de desarrollar innovaciones e investigaciones educativas en el mundo digital (Zapato-Ros, 2015). Así, Big Data en Educación se puede definir como la acumulación de datos masivos generados, creados y recreados en diversas instancias sociales, culturales, tecnológicas y educativas, cuya potencialidad pedagógica resulta ingente en términos de socialización y personalización del aprendizaje. Una de los elementos intrínsecos en su configuración interna como clave emergente dentro de la *nueva educación* es la interconectividad, entendida como el parámetro más didáctico dentro de la amplia y profusa ecología mediática. Nociones tales como producción, reproducción y distribución compartida de datos e informaciones educativas irán delimitando los nuevos caminos o vías de acción pedagógica (Domínguez, Álvarez & Gil-Jaurena, 2016).

Los efectos del Big Data están siendo cada vez más visibles en las cuentas de resultados de las grandes corporaciones tecnológicas, y su proyección en amplios sectores de transferencia mercantilizada de los datos personales a distintos niveles y dimensiones (Maroto, 2017). Pero, en el caso de la educación, pareciera que únicamente los gestores y administradores de los sistemas educativos están preocupados por la recopilación acrítica de datos sobre: rendimiento académico, necesidades personales de aprendizaje, dificultades o trastornos específicos de conducta, evaluaciones institucionales, evaluaciones de la calidad docente, etc. Ahora bien, la pregunta clave reside en por qué se recoge información sin articular procesos inteligentes, operativos y racionales de analítica de aprendizaje.

Big data y la analítica del aprendizaje deben entenderse como instrumentos pedagógicos relevantes que permiten aprovechar el potencial de enormes cantidades de datos que actualmente existen en formato digital, y que se pueden reconstruir durante los procesos de enseñanza y aprendizaje cuando se incorporan plataformas digitales a dicho proceso en la formación de los profesionales de la educación, por un lado, y en la atención educativa personalizada e inclusiva para el alumnado en general.

Existen múltiples objetivos en la aplicación y desarrollo práctico del Big Data en Educación, pero, en todo caso, resulta importante explotar las ventajas de las capacidades y estrategias didácticas actuales para el procesamiento inteligente de datos en grandes volúmenes y distintos formatos, para buscar su comprensión y mejorar la toma de decisiones en materia de proyección pedagógica inclusiva.

Con Big data y la analítica del aprendizaje es posible transformar grandes cantidades de datos educativos aparentemente inconexos en información de muy alta calidad para mejorar la formación del profesorado o, en su momento, mejorar la atención personalizada del alumnado con algún tipo de necesidad específica de apoyo educativo. Esta práctica supone un área de amplia repercusión en el desarrollo de distintas propuestas de educación inclusiva y personalizada, que puede ser aprovechada por docentes especialistas (profesorado de educación especial, audición y lenguaje, ATAL...) y profesionales de la orientación educativa, en otras palabras, quien conoce la información que tiene disponible en las bases educativas de datos (tipo Séneca, de otras instancias y agentes formativos como los CEP, y otros). Ni que decir tiene que para desarrollar una analítica de aprendizaje que sea propositiva, coherente y funcional, se requiere profesorado especializado y formado en esta tarea, o especialistas pedagogos o psicopedagogos en la gestión y procesamiento de datos educativos. Se requiere infraestructura de toda índole, pero con las posibilidades para el trabajo en la nube, y los recursos disponibles en los centros educativos, se puede partir de un primer nivel de concreción telemática en materia de hardware y software (Castro, 2015), y escalar a plataformas más sofisticadas y complejas cuando la combinación de hallazgos y descubrimientos en los datos y de los volúmenes a procesar lo justifiquen educativamente hablando. Sin lugar a duda, Big Data y analítica del aprendizaje son instrumentos emergentes que deben ser adecuadamente abordados desde enfoques educativos progresivos, inteligentes, sistemáticos, avanzados, creativos, innovadores y responsables.

*Progresivos,* en la medida en que estos datos deben ayudar a focalizar la mirada educativa más personalizada y especializada para diseñar, desarrollar y evaluar propuestas de inclusión educativa para alumnado con necesidades específicas de apoyo educativo.

*Inteligentes*, puesto que la analítica del aprendizaje derivada del Big Data debe servir de "soporte" institucional, de aplicación de acciones e iniciativas de mejora de las condiciones profesionales y pedagógicas de las instituciones educativas como organizaciones que aprenden, que se transforman para servir mejor a los intereses de todos los miembros de la comunidad escolar.

*Sistemáticos*, en relación a la recogida y análisis de datos mediante las técnicas estadísticas que soporten la complejidad de la realidad educativa y, ante todo, la rigurosidad al analizar la holística de la formación.

*Avanzados,* en tanto las derivadas y aprendizajes de la inmensa cantidad de datos recogidos en las bases de datos especializadas en educación va a permitir la mejora de los sistemas de recomendación didáctica, y, por supuesto, ayudar al establecimiento de nuevas guías y protocolos educativos más eficaces y eficientes en

el abordaje interdisciplinar de distintas modalidades o diversidades de corte situacional, contextual o personal.

*Creativos,* porque a pesar de lo que se pudiera pensar, no buscamos el control absoluto porque es imposible al no poderse medir toda la complejidad de la diversidad humana en los procesos educativos. Ahora bien, las respuestas educativas de los sistemas escolares del siglo XXI van a priorizar el pensamiento creativo, divergente y crítico al tener que conciliar distintas funciones de desarrollo humano en combinación con mayores y más complejos avances en la inteligencia artificial y la creciente digitalización de toda de decisiones en distintos ámbitos de la vida.

*Innovadores,* ya que se debe perseguir que las respuestas educativas a las necesidades específicas del alumnado y también de formación de los profesionales de la educación sean ajustadas y realmente útiles y funcionales para las personas y para las instituciones. La innovación requiere siempre equilibrio institucional y es que tan importante son las necesidades personales como sociales en el marco de organizaciones escolares que deben ser entendidas como comunidades de práctica o comunidades de aprendizaje.

*Responsables,* ya que los sistemas de recomendación didáctica no son recetas estandarizadas o acríticas, sino fundamentadas y concretadas en propuestas estudiadas y compartidas por los profesionales de la orientación y de la educación, atendiendo a los planteamientos de las familias del alumnado y sus propias necesidades. La responsabilidad implica una mejora sustancial de las prácticas docentes en términos de agilidad de las respuestas educativas, en tanto el Big Data pretende reconducir de forma eficaz la excesiva burocratización que existe hoy en día en las escuelas. No se trata de no recoger datos, sino de hacerlo de una forma más inteligente, colaborativa y, sobre todo, interconectada y en red lo cual permitirá un aprovechamiento óptimo de los recursos disponibles en materia pedagógica en distintos centros formativos tales como: escuelas infantiles, CEIP, IES, EOE, CEP, EOI, Conservatorios de Música, Escuelas de Danza, etc.

En las siguientes páginas, se va a realizar una revisión de los conceptos fundamentales vinculados con el Big data. Posteriormente se analizará el potencial del Big data, junto con el Data Mining, para la evaluación de programas específicamente la evaluación del impacto.

## 2. BIG DATA, MINERÍA DE DATOS, ANALÍTICA ACADÉMICO Y ANALÍTICA DEL APRENDIZAJE: CUESTIÓN DE TERMINOLOGÍA

Anteriormente, se ha rizado una aproximación al término Big data. actualmente, Big data hace referencia a la capacidad de la gestión de grandes cantidades de datos, así como a la profesión que ha surgido alrededor de esta tecnología; sin

embargo, el término Big Data está vinculado a otros que sería interesante diferenciar.

Así, tenemos Data mining, Academical analytic, o Learning analytic. A continuación se presenta una definición escueta de cada una de ellas, con la intención de tener claro un marco de trabajo con relación al Big data, y a las distintas estrategias que se han mencionado.

### 2.1. Minería de datos

En términos generales, el Data mining, o minería de datos en español, trata de desarrollar, analizar, y aplicar métodos informatizados para identificar patrones en grandes cantidades de datos. La aplicación de las técnicas de minería de datos al ámbito de la educación, está dando lugar a la aparición de la minería de datos educativa o educacional.

### 2.2. Analítica académica y analítica de aprendizaje

El concepto de analítica académica y de analítica de aprendizaje está muy relacionados. La analítica de aprendizaje, o Learning analytics en su expresión inglesa, sería el proceso de recogida de información educativa, su medición, su análisis, así como la presentación de los informes, de datos que se producen en el proceso de aprendizaje del alumnado y su contexto. La idea principal, es la de comprender ese proceso de aprendizaje.

Por su parte, la analítica académica consiste en la aplicación de herramientas de inteligencia de negocios para la toma de decisiones en las instituciones educativas. El objetivo de la analítica académica es ayudar a mejorar la institución educativa a través de la recogida de datos, la medición y la generación de informes de manera efectiva y operativa, donde se identifiquen los aspectos positivos así como las debilidades de la institución.

Ambas analíticas se fundamentan en el análisis de datos aplicado al Big Data.Por otro lado, aunque en ocasiones tanto la analítica académica como la de aprendizaje son identificadas con la minería de datos en educación, mantienen ciertas diferencias. Por ejemplo, en cuanto al origen, la minería de datos no es sólo anterior, sino que mantiene vinculación con el software educativo. Por su parte, las analíticas están más vinculadas a la web semántica.

Otra diferencia notoria es la finalidad. Según esto, la minería de datos se centra más en el uso de metodologías automatizadas o, en otras palabras, en el aprendizaje automático, que permite alcanzar una predicción eficiente de los procesos educativos; por otro lado, las analíticas están más centradas en comprender el proceso de enseñanza-aprendizaje, así como su vinculación con el contexto.

Al margen de cuestiones semánticas, desde aquí defendemos que, en la actualidad, la relación entre estas disciplinas son las siguientes: el Big Data constituye el conjunto de estrategias que permiten la recolección y gestión de datos de forma masiva; la minería de datos aporta las estrategias analíticas necesarias para identificar patrones; y finalmente las analíticas le dan un sentido práctico y orientan el uso de unas estrategias y otras para una finalidad determinada. Estas finalidades pueden ser la mejora de la institución educativa, o la comprensión del proceso de enseñanza y aprendizaje en el alumnado.

En los siguientes epígrafes, se va a presentar cómo el Big data y la minería de datos pueden ser utilizados para la evaluación de los programas y las políticas. De esta forma centramos el resto del capítulo en la analítica académica.

## 3. EVALUACIÓN DEL IMPACTO DE PROGRAMAS Y POLÍTICAS EDUCATIVAS

Cuando se habla de evaluación de programas, se hace referencia a un proceso de valoración, basado en procedimientos científicos, que tratan de dar respuestas a una serie de preguntas. Estas preguntas pueden ser de tipo descriptivo, normativo, o causales. Por ejemplo, se puede preguntar cómo se ha desarrollado un programa a lo largo de su trayectoria, o bien, si el programa ha utilizado los recursos de acuerdo a lo establecido. En el primer caso estaríamos ante una pregunta descriptiva, y en el segundo ante una cuestión normativa.

El tercer tipo de pregunta es causal. Un ejemplo de este tipo respondería a cuál es el efecto del programa en el entorno donde se ha desarrollado. Este tipo de preguntas son propias de la evaluación del impacto de programas, políticas o proyectos.

En cualquiera de los casos, la evaluación implica un proceso de monitorización. La monitorización consiste la recogida continua de información sobre la cual se soportará la aplicación de las estrategias analíticas correspondientes, y a partir de ahí se establecerán las conclusiones oportunas con relación al programa, el proyecto, o la política que se valoran.

La diferencia entre la evaluación en general, y la monitorización, consiste en que en la primera se llevará a cabo una toma de decisiones que afectará al conjunto o a parte de los elementos y agentes implicados en el programa que se evalúa. Según esto, la monitorización es solamente un recurso más para la evaluación. De hecho, la monitorización tiene sentido como elemento independiente; ya que, en muchas ocasiones, o bien no es necesario, o bien no se pretende una evaluación, sino un simple registro de toda la información que se genera a lo largo del programa. Pero siempre que se pretenda valorar algún aspecto del programa, por

ejemplo, la eficacia, eficiencia, o utilidad, se estará ante una evaluación, y no ante una simple monitorización.

En este sentido, la evaluación del impacto de un programa se debería llevar a cabo siempre que dicho programa fuese innovador, aplicable, fuese relevante para un determinado contexto, y no se hubiese valorado previamente; es decir, un programa, proyecto, o política, muy influyente para la toma de decisiones que afecten a otros programas. En otros casos, donde no se ven ninguna de estas condiciones, la evaluación del impacto se llevará a cabo en función de los intereses de aquellos que han diseñado el programa, proyecto, o política o en función de los intereses de los agentes implicados (stakeholders).

Un elemento básico para iniciar la relación del impacto, es la teoría del cambio del programa, proyecto o política. La teoría del cambio es una descripción de cómo los diseñadores del programa piensa que se va a producir la obtención de resultados a partir de las actividades propuestas. La teoría del cambio se basa en la cadena causal, a la que se añaden los supuestos teóricos, experimentales, o empíricos, que soportan y justifican la puesta en marcha de las diferentes actividades que implica el programa, proyecto, o política. Sin la teoría de cambio, el proceso de evaluación y, sobre todo, la valoración del impacto del programa, es bastante limitado. La teoría del cambio se puede representar de múltiples formas; la habitual es como una descripción narrativa de la misma, pero también se puede recurrir al marco lógico, a la representación gráfica, o a la simple cadena causal añadiendo comentarios sobre los supuestos teóricos que sustentan la puesta en marcha de las distintas actividades.

Al llevar a cabo la evaluación de un programa, es necesario analizar el contexto dónde se realiza el programa, y determinar junto con otros factores, cuál es el mejor plan de actuación. Estos planes de actuación son los llamados diseños de evaluación. En caso de la evaluación de programas, y especialmente elaboración del impacto de los programas, es posible identificar tres escenarios generales con relación a los diseños:

El primer escenario lo constituyen aquellos diseños donde se pueden llevar a cabo ensayos controlados aleatorios. Estos diseños son aquellos donde es posible seleccionar a los sujetos que van a ser evaluados, así como asignarlos a las condiciones o niveles que determina el programa, incluyendo una condición en la cual no se lleva a cabo ese programa. Estos diseños, proceden directamente de la disciplina de métodos de investigación y se identifican con los diseños de investigación experimental. Son diseños difícilmente aplicables de forma generalizada, puesto que en educación es complicado poder llegar a tal grado de control de manera que se pueda seleccionar a los sujetos y asignarlos a las diferentes condiciones experimentales. Sin embargo, son diseños que algunas instituciones internacionales exigen para garantizar la validez de la evaluación. Tanto es así que a este tipo de

diseños se les suele conocer como "la regla de oro de la evaluación" (e.G. IPA: https://www.Poverty-action.Org/about/randomized-control-trials).

El segundo escenario es aquel donde solamente es posible un control parcial de las variables. Esto implica que, o bien no se pueden seleccionar aleatoriamente a los sujetos que participan en la evaluación, o bien no se pueden asignar a los diferentes niveles del programa de forma aleatoria. Aunque, sobre el papel, estos diseños presentan la misma estructura que los diseños totalmente aleatorios, tienen la desventaja de que el control no es total, y por lo tanto sus resultados siempre pueden ser sometidos a un juicio de valor. Estos diseños son bastante habituales en el ámbito educativo, pero no por ello sus resultados deben tomarse como totalmente ciertos y válidos; es muy importante que sean valorados de manera crítica, tratando de buscar vías alternativas que los apoyen. En términos de metodología de investigación, estos planes suelen llamarse como diseños cuasi experimentales.

El tercer escenario, lo constituyen todos aquellos donde no es posible ningún control sobre las elementos que integran la evaluación. En este caso, es necesario buscar formas alternativas de evaluación que no coinciden ni con los diseños experimentales, ni con los diseños cuasi experimentales. Evidentemente, los resultados que se obtienen de estos procesos son discutibles, y por ello tienden a integrar diferentes estrategias y metodologías, formando una especie de triangulación metodológica para garantizar la validez de sus resultados hasta cierto punto. A pesar de su debilidad en cuanto a la validez de sus resultados, son los diseños que más surgen y aparecen en la evaluación en Ciencias Sociales. Por tal razón este capítulo se centra en ellos, y en cómo el Big Data y la minería de datos pueden ayudar a los mismos. A este tipo de diseño se le puede llamar "de coherencia" puesto que tratan de analizar la coherencia que existe entre la teoría del cambio, y los resultados tanto parciales, como finales, que se obtienen con los programas.

## 4. ESTUDIOS DE COHERENCIA

Como se ha dicho anteriormente, los estudios de coherencia son aquellos que se llevan a cabo cuando es imposible poner en marcha un diseño experimental o un diseño cuasi experimental. Aunque son discutibles en cuanto a su potencia metodológica, constituyen casi la única alternativa viable cuando es imposible llevar a cabo otro diseño más potente.

Los diseños de coherencia se centran en buscar y recopilar información útil, que permita analizar la verosimilitud de la atribución causal que se establece en las teorías del cambio, entre sus actividades y los fines que persiguen. Para ello recurre a todos los procedimientos metodológicos y estrategias que le sean útiles.

De tal forma que, cuando es posible, se contrastan sus resultados parciales con relación a los indicadores previstos, a través de técnicas estadísticas, técnicas cualitativas, o cualquier otro tipo de estrategia.

Por tanto, los estudios de coherencia se basan fuertemente en la cadena causal que esté implícita en la teoría del cambio del programa. Es indispensable tener claro cuáles van a ser los resultados parciales que se obtendrán a lo largo del desarrollo del programa, y que deben estar claramente recogidos en la teoría del cambio. De esta forma es posible ir comprobando la sucesión de eventos a medida que el programa va desarrollándose. El equipo de evaluación deberá ir analizando dichos resultados parciales, contrastándolos con lo que estaba previsto en la teoría del cambio, y valorar si coinciden, hasta qué grado coinciden y, si no es así, qué aspecto del programa están fallando.

Como se ha dicho anteriormente, para el desarrollo de estos diseños, se utiliza cualquier procedimiento científico que sea útil. Entre otros, se recurre a los siguientes:

- Análisis de la coherencia interna de la teoría del cambio. Se trata de valorar si la teoría del cambio sigue un argumento lógico y se basa en evidencias. Lo suele hacer el equipo de evaluación.
- Predicción de los resultados intermedios. Consiste en ir comprobando que se van generando los resultados previstos en la teoría del cambio a lo largo del programa, no solamente los resultados finales, sino también aquellos resultados o impactos parciales del mismo.
- Contraste de los resultados con predicciones de expertos.Se consulta a un grupo de expertos para que aporten sus opiniones con relación a la teoría del cambio. Específicamente se les pide dos cuestiones, la primera que valoren la lógica interna de la cadena causal, y la segunda que realicen un esfuerzo predictivo sobre qué resultados se obtendrán y de qué naturaleza serán los mismos.
- Entrevistas a participantes clave. Una vez realizado el programa, o bien una vez que se van obteniendo resultados parciales, se pueden hacer entrevistas a participantes clave del programa con la intención que aporten su perspectiva sobre mismo. Al igual que antes, el objetivo es que realicen una valoración sobre los resultados que se van obteniendo con relación a la teoría del cambio propuesta.
- Descartar hipótesis alternativas. Es posible, que la realización de un programa coincida con otros eventos que afecten a su resultado. En tal caso, el efecto del programa será resultado de su capacidad más la interacción con dichos eventos. Este aspecto de interacción es muy estudiado por la metodología de investigación en Ciencias Sociales. En ocasiones la interacción puede ser contraproducente para el efecto del programa, bien porque vaya en contra de sus objetivos, o bien porque, en ocasiones, puede mejorar el efecto del mismo. En cualquier caso, los estímulos del contexto se entienden como variables extrañas, o como variables

mediadoras, que deben ser controladas. Sin embargo en este tercer contexto, no se pueden llevar a cabo diseños que controlen el efecto de variables.Pero no todo está perdido, puesto que se pueden llevar a cabo distintas estrategias de investigación social para identificar posibles explicaciones alternativas a los resultados que se van obteniendo con el programa. Sería el caso de un grupo focal que vaya analizando el contexto donde se desarrolla y, a modo de radar, trate de identificar hipótesis explicativas alternativas al programa. Si no se encontrase ninguna otra explicación más plausible a los cambios registrados, se supondría que el programa es la causa principal de dichos cambios. Esta estrategia, es típica del Análisis de la Contribución (Mayne, 2008).

Todas estas estrategias van destinadas a evitar todos los problemas metodológicos que van asociados con estos diseños. A pesar de todo, la validez interna de estos diseños siempre está bajo sospecha. No obstante, son muchas las ocasiones donde no es posible desarrollar otro tipo de diseño más potente, siendo los diseños de coherencia la mejor alternativa disponible. En este sentido, el trabajar con datos masivos, y aplicar estrategias de minería de datos permite que, de forma empírica, los resultados que se van obteniendo en un programa, sean susceptibles de ser plausibles. Esto implica tener una gran cantidad de datos, y que los resultados que se vayan obteniendo sean coherentes con la teoría del cambio propuesta. Aquí radica la importancia de tener una teoría del cambio bien fundamentada en cada programa que se evalúe.

## 5. UN EJEMPLO: BIG DATA COMO FUENTE DE DATOS Y MINERÍA DE DATOS COMO ESTRATEGIA DE ANÁLISIS

Para este ejemplo se ha utilizado una matriz de datos simulados de un supuesto curso realizado en internet a través de la plataforma Moodle. El número de supuestos alumnos es de 1792. A partir de ahora se le denominará “entidades” o “instancias” puesto que es la nomenclatura habitual de referirse a las filas de una matriz de datos en Minería de Datos. Por su parte, se han generado 20 variables, que representan los registros temporales que genera la plataforma Moodle en sus registros. A estas variables las llamaremos “atributos”.

En primer lugar se desarrolla la parte propia del Big Data. Esta parta corresponde a los administradores del sistema Moodle. Son estos los que ponen en marcha el software y hardware necesario para registrar, almacenar y gestionar todos los registros de Moodle. Una arquitectura Big Data incluye los siguientes componentes (adaptado de Soares, 2012):

- Fuentes de Big Data.
- Analítica de flujo continuo.
- Bases de datos.
- Integración de Big Data.
- Analítica de textos.
- Descubrimiento de Big Data.
- Calidad de Big Data.
- Metadatos.
- Gestión de políticas de información.
- Gestión de datos maestros.
- Data warehouses y data marts.
- Analítica y reporting de Big Data.
- Seguridad y privacidad de Big Data.
- Gestión de ciclo de vida de Big Data.
- La nube.

Una vez concluido el curso se recuperan los registros. Moodle puede generar dos informes, el de registros y las calificaciones. El informe de registros es sólo un listado de eventos listado por fechas. Para ser útil para el análisis, es necesario pasarlo a una matriz, donde cada columna sea una fecha o un evento, y en cada fila se registre el número de veces que un alumno (entidad) accedió a la plataforma o a un recurso en concreto. Ambos registros se unen, formando una sola matriz de datos que pasará al proceso de análisis.

Para este ejemplo se ha elegido la aplicación ORANGE versión 3.18 (Demser et al. 2013). Orange es un software multiplataforma de Minería de Datos desarrollado en Python. Tiene un interface gráfico muy intuitivo y dispone de una gran cantidad de material documental de apoyo. A fecha de enero de 2019, este software está disponible de forma gratuita en el sitio web https://orange.biolab.si/.

El procedimiento seguido en este ejemplo se ajusta al proceso conocido como Descubrimiento de Conocimiento en Bases de datos o KDD por su nomenclatura en Inglés (Knowledge Discovery in Databases). El KDD es simplemente una propuesta de pasos a seguir en Minería de Datos para encontrar un modelo o representación válida y útil de los datos, que describa posibles patrones subyacentes de acuerdo a la información. Estos pasos son: selección de datos, pre-procesamiento que es preparar (se ha comentado anteriormente con relación a la matriz de datos) y limpiar los datos, transformar los datos de acuerdo a la naturaleza de los mismos, la fase propiamente de Minería, y finalmente la interpretación del modelo y su evaluación.

Tras tener los datos preparados en una sola matriz, se revisó para garantizar que no había errores de registro. En esta ocasión no se generaron variables deri-

vadas, solamente se procedió a identificar la calificación final del curso como atributo clase. El atributo clase es aquel que se utiliza por algunos modelos como variable a predecir. Esto es necesario solamente cuando se utilizan modelos de aprendizaje supervisado.

A continuación, se procedió al análisis. El primero de ellos fue la realización de un análisis clúster jerárquico que permite identificar grupos. El procedimiento puede seguirse en el vídeo https://youtu.be/dJ5z2SRwzgs en YouTube, realizado por el equipo de Orange. En nuestro caso, los resultados muestran la existencia de tres grupos a un ratio del 90%. El primero muestra puntuaciones entre 88.80 y 91.20. El segundo grupo lo forma alumnado con 90.09 puntuaciones solamente y el tercero presenta puntuaciones tanto bajas (e.g. 84.27) como muy altas (e.g. 97.61). En la figura 1 se observa la estructura clúster indicada.


**Figura 1.-** Análisis clúster basado en promedio, con 5 niveles de profundidad y un ratio del 90%.

Las características descriptivas de cada clúster se puede observar en la figura 2. En dicha imagen se puede observar que el clúster 1 es mucho más homogéneo que el clúster 3, aunque es el clúster 2 quien está compuesta con alumnado que tiene la misma puntuación.

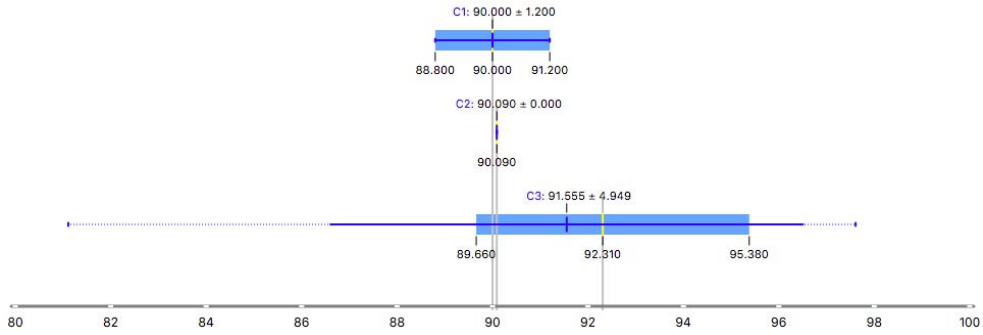


**Figura 2.-** Boxlplot de los clúster de la figura 1.

Para indagar un poco más, se ha cambiado el nivel de agrupación, situándolo en 80% (figura 3).

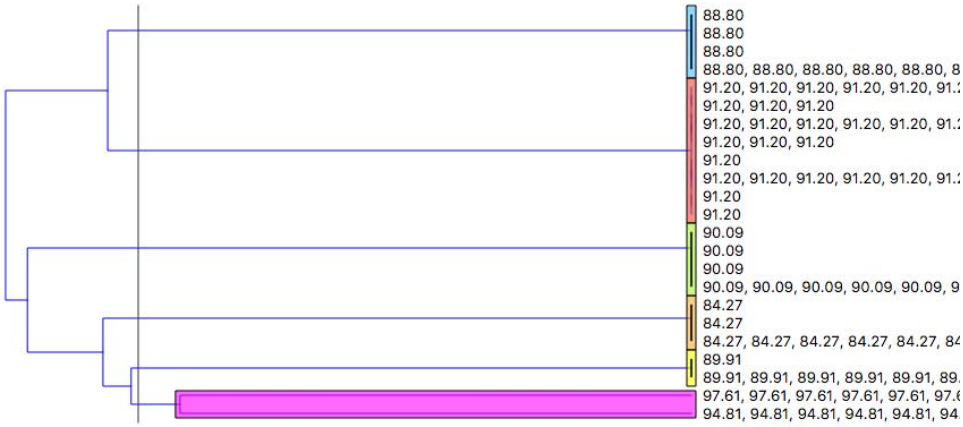


**Figura 3.-** Análisis clúster basado en promedio, con 5 niveles de profundidad y un ratio del 90%.

En este caso, se puede observar en el análisis visual de boxplot (figura 4) que solamente el último clúster agrupa alumnado con variabilidad de puntuaciones, mientras que el resto tienen las puntuaciones finales constantes.

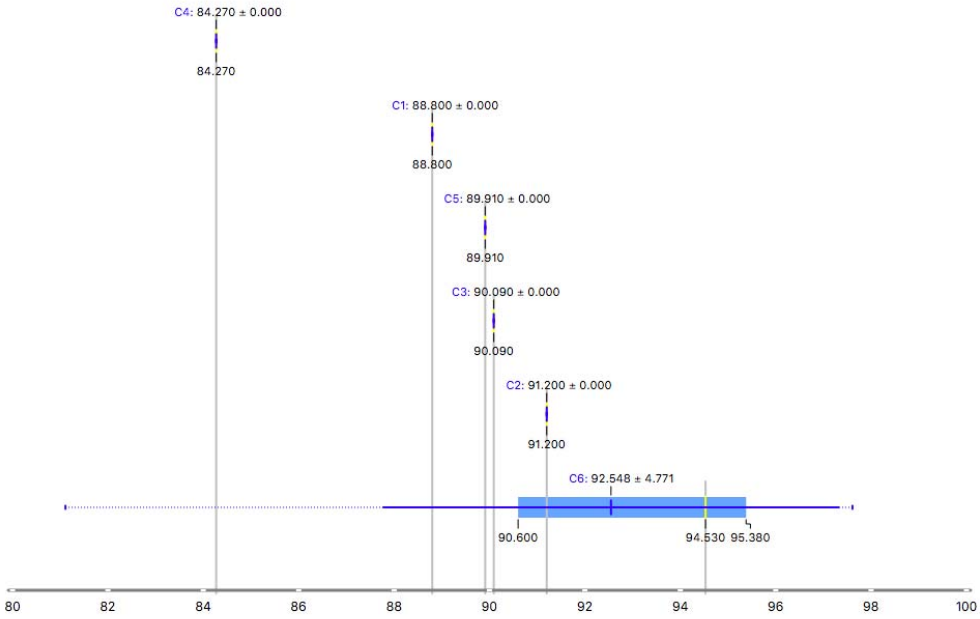


**Figura 4.-** Boxlplot de los clúster de la figura 3.

Siguiendo con la indagación sobre la estructura de los datos, se procede a realizar un análisis de clasificación aplicando el análisis árbol de decisión. En la figura 5 se puede ver que la variable que más peso tiene en la clasificación es S2016-nov-45. Si un sujeto tiene más de 11 puntos, el siguiente atributo que más pesa para clasificarlo es S2017-ene-04. El resto de alumnos se clasificarán a partir del atributo S2016-dic-48. De esta forma, el árbol va indicando que variables son las que más importancia tienen para clasificar las instancias en función de sus puntuaciones.

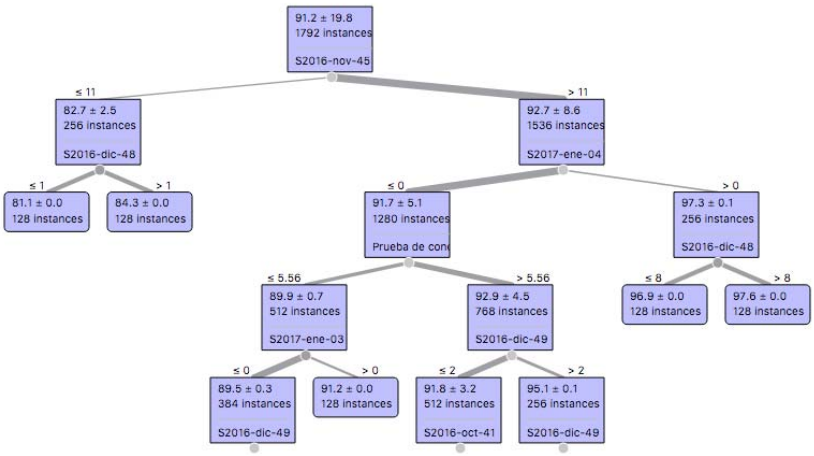


**Figura 5.-** Representación del árbol de decisión.

A través de este procedimiento sabemos, por ejemplo, que el alumnado que accedió 11 veces o menos al curso en la semana 45 del año 2016 (eso es lo que indica el nombre del atributo S2016-nov-45) obtendrá un 81.1 en la calificación final si sólo accedió una vez o menos en la semana 48 de dicho año. Si esa semana accedió más de una vez, su puntuación será de 84.3. Por tanto, se puede suponer que las personas que acceden al contenido esa semana obtendrán casi tres puntos más que aquellos que no lo hacen.

En el lado opuesto, el árbol nos indica que las puntuaciones más altas las obtuvieron aquellos alumnos que accedieron más de 11 veces la semana 45 de 2016, que además accedieron alguna vez la semana cuarta de 2017 y que también habían accedido más de 8 veces la semana 48 de 2016. Este alumnado obtuvo una puntuación de 97.6 puntos.

La última fase sería la evaluación del modelo. Para ello habría que valorar el error de generalización y el error de clasificación. El primero consiste en analizar si los modelos pueden utilizarse para contextos similares a los originales de donde proceden los datos. Para ello se puede dividir el conjunto de datos en dos grupos, uno de ellos se utilizará para generar los modelos y el otro para probarlo.

El segundo consiste en ver cuantas entidades se clasifican correctamente. Se puede comprobar simplemente viendo el porcentaje de entidades que han sido clasificadas correctamente en los mismos datos originales.

En el este vídeo de YouTube: https://youtu.be/pYXOF0jziGM;se puede ver cómo realizar el análisis de la bondad del modelo en Orange. Este programa permite también comprobar la capacidad de distintos algoritmos sobre los mismos datos. En el caso de los datos utilizados en el ejemplo, los resultados mostraron un

nulo ajuste, lo que era de esperar puesto que eran datos simulados sin estructura aparente.

En la figura 6 se muestra el flujo de conocimiento utilizado en Orange de forma gráfica.

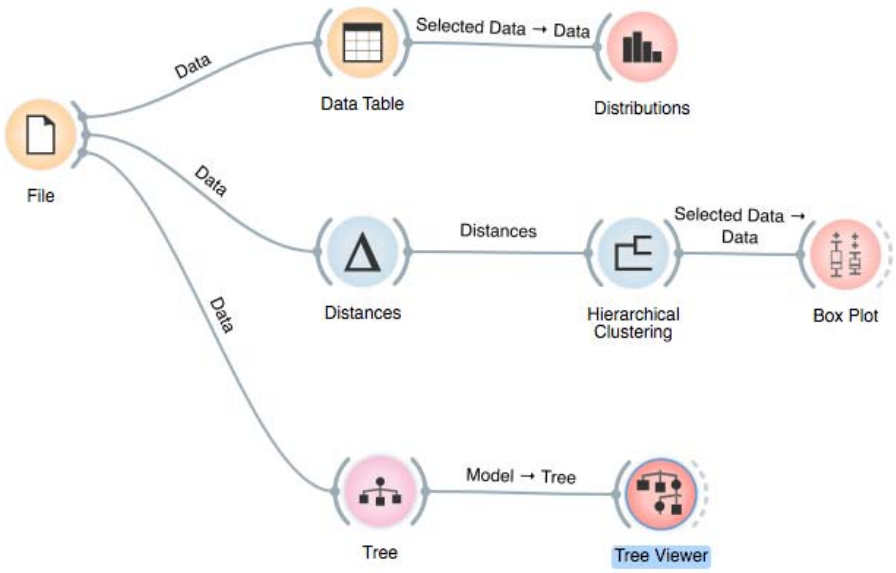


**Figura 6.-** Flujo de conocimiento utilizado para el ejemplo.

## 6. CONCLUSIONES

Vivimos un tiempo acelerado de cambios en el mundo de la formación. La preocupación por la implementación progresiva de medidas y acciones de garantía de calidad se vincula con la preocupación porque la inversión en educación sea rentable en términos de operatividad, eficiencia y también personalización del aprendizaje y rentabilidad social.

La aplicación de las nuevas herramientas de minería de datos y el Big Data en educación se configura como una excelente oportunidad para mejorar los sistemas educativos y las estrategias didácticas a fin de predecir las posibles repercusiones e impacto de medidas de innovación pedagógica, a la vez que mejorar de forma sustancial el rigor y coherencia en los análisis de los aprendizajes de los distintos agentes que conforman las comunidades escolares, sociales y científicas.

## 7. REFERENCIAS

Castro, P.R. (2015). Paradigmas analíticos en entornos virtuales y de aprendizaje: una revisión de sus principales puntos de encuentros y diferenciaciones teóricas y de enfoque. *Revista educación y tecnología*, 7, 91-106.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., y Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* ,14, 2349–2353.

Domínguez, D., Álvarez, J.F. y Gil-Jaurena, I. (2016). Analítica del aprendizaje y Big Data: heurísticas y marcos interpretativos. *Dilemata*, 22, 87-103.

Euronews (en español). (2015, 22 mayo). *Big Data al servicio de la educación - learning world* [Archivo de vídeo]. Recuperado de: https://www.youtube.com/watch?v=w07yJkHbIwE

García-Aretio, L. (2017). Educación a distancia y virtual: calidad, disrupción, aprendizajes adaptativo y móvil. *RIED. Revista Iberoamericana de Educación a Distancia*, *20*(2), 9-25.

Maroto, C. (2017). Big Data y su impacto en el sector público. *Harvard Deusto Business Review*, 16-25.

Mayne, J.(2008). *Contribution Analysis: An approach to exploring cause and effect, ILAC methodological brief.* Recuperado de: http://www.cgiar-ilac.org/files/ILAC_Brief16_Contribution_Analysis_0.pdf

Serrano-Cobos, J. (2014). Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos. *El profesional de la información*, *23*(6), 561-565.

Soares, S. (2012). *Big Data Governance: An Emergin Imperative*. Boise, Idaho, USA: MC Press.

Zapata-Ros, M. (2015). Analítica de aprendizaje y personalización. *Campus virtuales*, *2*(2), 88-118.