

DATA SCIENCE COMO CATALIZADOR DE MEJORAS EN EDUCACIÓN

Sebastián Flores, U-Planner, sebastian.flores@u-planner.com
Rodrigo Hernández, U-Planner, rodrigo.hernandez@u-planner.com
Cristián Espinoza, U-Planner, cristian.espinoza@u-planner.com

RESUMEN

Data Science permite convertir datos en conocimiento aplicable a la toma de decisiones. Esta filosofía de acción ha sido ampliamente replicada en múltiples industrias, entre las que destacan particularmente las de Tecnologías de la Información. La aplicación de data science en educación presenta variados desafíos, pero posee un gran potencial para optimizar la eficiencia de sus procesos internos y recursos, y consecuentemente mejorar la experiencia educativa de los estudiantes. En el presente documento, se caracterizan los tipos de datos disponibles para Institución de Educación Superior, para dar discusión a dos aplicaciones de data science a la educación: planificación académica y deserción estudiantil.

PALABRAS CLAVES: Data Science, Educación, Algoritmo, Machine Learning, Optimización, Planificación Académica, Retención Estudiantil.

INTRODUCCIÓN

La revista Harvard Business Review llamó a data scientist la “carrera más atractiva del siglo 21” el año 2012. Más recientemente, la empresa de búsquedas e información de trabajo Glassdoor nombró esta carrera como el “mejor trabajo del año” en el 2016. Si bien Data Science (ciencia de datos) y sus practicantes, data scientists (analistas de datos) resultan términos que se han vuelto del dominio habitual, muchas veces son utilizados con cierta confusión. Data science puede entenderse como una ciencia interdisciplinaria que toma herramientas de múltiples dominios con la finalidad de explorar y analizar algún conjunto de datos con la finalidad de proporcionar información para la gestión de acciones. En resumen, data science puede entenderse como convertir datos en conocimiento para la toma de decisiones. Tareas habituales de un data scientist consisten en hallar fuentes apropiadas de información, extraer y transformar datos, lograr una comprensión profunda de los datos y del proceso a modelar, crear y calibrar modelos determinista-estadísticos típicamente provenientes de Inteligencia Artificial y Machine Learning, utilizar dichos modelos para el análisis y/o predicción de sucesos futuros y visualizar grandes conjuntos de datos. De todas las tareas anteriores, la extracción y transformación de datos constituyen habitualmente la mayor parte del tiempo, pues resulta la tarea que es más difícil de automatizar y requiere conocimiento del origen de los datos y de la finalidad del modelamiento. A pesar que el término data science aparece en la última década, las tareas realizadas por un data scientist preceden a la definición del término. En efecto, en el artículo “From Data Mining to Knowledge Discovery in Databases” de 1996 ya se mencionaba las etapas y desafíos que caracterizan a Data Science. La gran diferencia que ha surgido en los últimos años ha sido la maduración de un sinnúmero de lenguajes y librerías que facilitan la transformación, procesamiento, modelamiento y visualización de los datos. Adicionalmente, se han vuelto conocidos casos de empresas como Facebook, Netflix o Amazon, que sustentan gran parte de su éxito en el conocimiento que pueden rentabilizar del comportamiento de sus usuarios. Estamos sin lugar a dudas en una época en la cual es posible obtener un gran volumen de información, y en la cual la capacidad de procesar dicha información y extraer conocimiento valioso es el factor diferenciador entre las empresas. La aplicación de data science en educación se encuentra todavía en sus etapas iniciales y presenta un sinnúmero de desafíos, entre los cuales destacan

el acceder a fuentes de información suficientemente completas y contar con expertos trabajando en el área. Existe un desafío adicional puesto que los datos de la institución contienen información confidencial y que requieren un tratamiento especial para garantizar el procesamiento seguro de la información, además de restringir los accesos a los datos que pudieran resultar más sensibles.

EJEMPLOS DE APLICACIÓN DE DATA SCIENCE

En minería existe un gran potencial de aplicación de data science, entre los cuales destacan el mejorar la productividad, eficiencia y seguridad utilizando los datos registrados por sensores y cámaras. Un caso de interés lo constituye el concurso “Gold Rush Challenge”, en el cual la empresa Integra dispuso 6 terabytes de datos correspondientes a 75 años de exploración minera en un concurso online de competición de equipos de data science. El equipo ganador combinó técnicas de machine learning y métodos tradicionales en geología, prediciendo lugares con gran potencial de alto contenido de oro que habrían sido descartados utilizando únicamente los métodos tradicionales. En el sector bancario, existe un creciente interés por la aplicación de data science debido al gran volumen de información registrado por cada cliente respecto a sus diversas transacciones, y a partir de las cuales es posible generar inteligencia de negocio como detección de fraude, credit scoring, o estimación de riesgo, o bien acciones proactivas como ofrecimiento de créditos gatillados por comportamiento de compra. Un ejemplo concreto lo constituye el Banco Goldman Sachs, que tiene un área de Data Science & Engineering completamente dedicado a la generación de herramientas analíticas específicas a dicha institución financiera. Sin embargo, la mayor parte de las empresas que actualmente utilizan data science como parte fundamental de su modelo de negocios lo constituyen empresas de tecnologías de la información, como Facebook, LinkedIn, Twitter, Google, Quora, Groupon, Netflix, AirBnB o Uber, entre muchas otras. Estas empresas tienen departamentos dedicados al análisis y generación de nuevos productos basados en la información que recolectan de sus usuarios, permitiendo crear una mejor segmentación de los clientes en función de sus gustos y comportamiento, y diseñando acciones específicas para cada uno de estos segmentos de mercado.

APLICACIÓN DE DATA SCIENCE EN EDUCACIÓN

La educación superior se encuentra inserta en un contexto de grandes números y alta complejidad. Por una parte se trata de, potencialmente, decenas de miles de estudiantes participando en un alto número de programas académicos, donde cada uno de estos jóvenes posee características psicológicas, de origen y de preparación académica diferentes. Puesto que el perfil de egreso común para una carrera debe obtenerse a pesar de perfiles de ingreso variables y disímiles, se requiere que la actividad académica que mueve al alumno hasta la titulación y egreso esté diseñado de forma que se adapte a estas circunstancias. El gran número de estudiantes implica a su vez coordinar y movilizar grandes conjuntos de personas bajo recursos limitados, debiendo medirse el avance curricular de cada uno y la mejor forma de apoyo en el proceso de formación y aprendizaje.

Modelar una institución es una tarea compleja puesto que ésta puede tener múltiples campus repartidos en una o más ciudades, con reglas complejas restringiendo la toma de ramos entre ellos, y manteniendo de manera simultánea diversas jornadas (regular, vespertina, sabatina, etc.) y distintos períodos académicos (semestres, trimestres o cuatrimestres). A continuación se detalla a modo de ejemplo un conjunto mínimo de datos a considerar para modelar una Instituciones de Educación Superior:

Datos de Docencia:

- Carreras y planes de estudio activos.
- Cursos, definidos en términos de contenidos, evaluaciones, créditos y prerrequisitos.
- Datos de Infraestructura:
 - Campus definidos en términos de los edificios que lo componen.
 - Definición de módulos horarios.
 - Tipo de períodos académicos y jornadas.
 - Sala con atribuciones de tipo de uso (clase, laboratorio, gimnasio, etc.), capacidad y ubicación.
- Datos de los Docentes:
 - Datos de contacto y títulos del docente.
 - Tipo de contrato y sueldo.
 - Historial de cursos que ha dictado, con resultados de aprobación y evaluación docente.
 - Disponibilidad espacio-temporal.
- Datos de los Estudiantes:
 - Datos de contacto y casos especiales.
 - Historial de asignaturas con notas finales y parciales.
 - Historial por estudiante, indicando los periodos académicos y el estado académico del alumno.

Adicionalmente, es posible que se cuente con información previa al ingreso del estudiante a la institución, como notas en educación media, resultados de pruebas de selección o preferencia de postulación; con información familiar como la estado civil, número de hermanos, nivel educacional de los padres, situación laboral del estudiante y sus padres, y/o situación socioeconómica. También resulta habitual almacenar información respectiva a los datos de financiamiento y el historial de pagos del estudiante, aunque esta información puede resultar de difícil acceso por ser sensible o no estar disponible digitalmente. En algunos casos es también posible que se cuente con información de resolución temporal muy fina, como la asistencia a clases, notas parciales, uso de biblioteca o recursos online, situación que sin duda se verá acrecentada en el futuro con el uso de dispositivos inteligentes y conectados a la llamada Internet of Things.

El gran volumen y variedad de los datos, y la alta complejidad de estos, inciden en que en la actualidad muchas decisiones en la educación superior sean tomadas ya sea de forma intuitiva o en base a las soluciones definidas para períodos anteriores, debido a que la complejidad del problema la lleva a recurrir a estas aproximaciones de solución. Es aquí donde la innovación de tomar técnicas de data science para modelar distintos aspectos del proceso académico resulta crucial para mejorar los resultados de aprendizaje, optimizar el uso de los recursos (y por tanto aumentar la cobertura potencial) y llevar la operación universitaria a un proceso de personalización de la experiencia de aprendizaje.

En la experiencia de U-Planner, las instituciones de educación superior almacenan la información anterior de manera parcelada en distintos sistemas, situación que algunos autores denominan silos de información. Esto muchas veces impide a la institución una visión completa de la información, facilitando la existencia de duplicidad de datos y dificultando el mantenimiento de éstos. Los sistemas de almacenamiento de información de las IES han sido tradicionalmente realizados para respaldar y almacenar la información, y no para que ésta sea utilizada habitualmente por la universidad para optimizar sus procesos. Otra situación que dificulta la utilización de los datos en los modelos lo constituye la incompletitud de la información, puesto que algunos de los datos han sido almacenados solamente a partir de cierto año o existen años en los que la información no fue registrada de manera apropiada.

U-Planner ha detectado diversas áreas en las cuales las IES requieren optimización de sus procesos, y para las cuales posee diversos productos, como se muestra en la Figura 1.



Figura N°1. Áreas estratégicas para una Institución de Educación Superior, y productos U-Planner que permiten solucionar algún aspecto de ellas.

Un primer ejemplo de aplicación de data science a la educación lo constituye la planificación académica, que se entiende como la asignación de horarios y salas de clases de manera que se respeten una serie de reglas de negocio. Ejemplos de estas reglas son, por ejemplo, que los alumnos tengan poca o nula incompatibilidad horaria, que se respeten restricciones de infraestructura, y que exista una asignación de preferencias de ciertas materias en algunos horarios, entre muchas otras. Este problema, conocido como *course scheduling* en inglés, es de tipo NP-completo y por tanto su optimización es NP-hard, que corresponde a la familia de problemas más difíciles de resolver en computación. En la práctica, esto se traduce en que hallar la solución óptima para la planificación académica de una institución utilizando software de optimización tradicional requiera un tiempo de cálculo que puede ser prohibitivamente grande. Sin embargo, la planificación académica de una institución debe obtenerse en un tiempo breve,

puesto que la planificación obtenida requiere ser validada por múltiples agentes, que frecuentemente realizan modificaciones que alejan la solución del valor óptimo y que corresponden a restricciones adicionales que no fueron consideradas inicialmente. Un mejor procedimiento consistiría en preparar múltiples planificaciones de manera simultánea para distintos conjuntos de restricciones y luego seleccionar aquella que se considere mejor bajo una serie de métricas. Lo anterior se logra cuando, en lugar de buscar el óptimo global, el conocimiento que se posee del problema y de las características de una “buena” solución se traducen en la implementación de diversas heurísticas que permiten encontrar soluciones cercanas al valor óptimo en menor tiempo. Este problema resulta interesante como aplicación de data science puesto que la resolución temporal tanto de los datos de entrada como de salida son relativos al periodo académico a predecir, lo que se traduce en que el volumen de datos requeridos para dar solución al problema sea pequeño y fácil de generar por parte de la institución. U-Planner ha desarrollado un producto de software, U-Planning, que mediante una interface web permite de manera sencilla ejecutar un cálculo de planificación académica. Casos de interés lo constituyen consultorías realizadas a diversas instituciones utilizando este software, que permiten obtener planificaciones académicas comparables a través de una serie de mediciones de interés como la movilidad de los alumnos, el porcentaje de uso de silla y el porcentaje de uso de salas. En una de estas consultorías, una institución de educación superior en Latinoamérica obtuvo como resultado la posibilidad de reducir el número de edificios requeridos para cumplir con la programación académica.

Un segundo ejemplo de aplicación de data science en instituciones de educación superior podemos encontrarlo en la predicción de deserción estudiantil. La deserción se define como el abandono prematuro de un programa de estudios antes de cumplir con los requerimientos del título o grado. Las Instituciones de Educación Superior en Latinoamérica comparten tasas de deserción que pueden llegar al 70%, lo cual puede impactar en la viabilidad del proyecto educativo debido a la merma en los ingresos que provoca la disminución de los estudiantes por cohorte y tiene además un importante efecto en el prestigio y acreditación de la institución. El análisis de la deserción estudiantil ha sido investigado por múltiples autores, entre los cuales destacan Spady, Tinto y Ethington. Los modelos de estos autores coinciden en explicar la deserción a partir de una serie de factores, que pueden ser mayormente sociológicos (Spady, 1970), integración académica y social (Tinto, 1975) o conductas de logro (Ethington, 1990). Mediante técnicas de data science es posible convertir la información disponible en factores de deserción, calibrar el modelo que mejor se aplica a estos datos, y utilizarlo para predecir el comportamiento de deserción de los alumnos que todavía se encuentran en la institución. La primera etapa, conversión de datos en factores, resulta crucial puesto que los datos disponibles en la universidad no son directamente utilizables debido a que pueden contener información errónea o incompleta, y muchas veces requieren una transformación para ser útiles. Por ejemplo, la fecha de nacimiento no puede utilizarse directamente para estudiar la deserción de un estudiante, pero al utilizarla para obtener la edad del estudiante en los distintos semestres puede utilizarse como una medida de la madurez del alumno. Similarmente, las notas del estudiante en sus cursos no son directamente aplicables ni comparables entre alumnos de distintas carreras, pero al crear factores de deserción como porcentaje de completitud de su plan académico y rendimiento académico promedio, se puede explicar de mejor manera la relación que existen entre las notas de un estudiante y la deserción. Los factores pueden ser individuales (edad, género, estado civil, procedencia, etc.), académicos (tipo de colegio de procedencia, notas de estudios secundarios, puntaje en pruebas de selección o ingreso, porcentaje de créditos aprobados, promedio de notas en primer año, asistencia a clases, etc.), socioeconómicos (situación laboral de estudiante y padres, nivel educativo padres, quintil de grupo familiar) u otros dependientes de la realidad de la institución y el país. En segundo lugar, la calibración del modelo es una de las etapas más complejas y que requiere una programación extremadamente flexible y dinámica. Esto, pues se

requiere que el modelo de retención pueda funcionar a partir de los factores de deserción con los cuales cuenta la institución, que pueden variar en número y calidad de una institución a otra, e incluso dentro del tiempo para una misma institución. La calibración debe realizarse cada vez que se actualiza o genera nueva información, de manera que el modelo predictivo sea en todo momento lo más preciso posible. Los modelos a calibrar corresponden a la familia de algoritmos de clasificación de aprendizaje supervisado, ampliamente conocidos en Machine Learning, y donde el algoritmo específico a utilizar debe ser elegido en función de su habilidad para efectuar buenas predicciones y permitir la interpretabilidad de los parámetros del modelo. Por último, la etapa de predicción permite aplicar el modelo calibrado a los datos de estudiantes activos de la institución, para obtener una estimación de la probabilidad de que el alumno deserte en el semestre en curso y a lo largo de su carrera. Resulta importante que todo el proceso anterior sea encapsulado dentro de un software que sea de fácil utilización para la persona que finalmente usará la predicción para realizar contacto con alumnos en riesgo, normalmente perteneciente al equipo de Bienestar Estudiantil o su equivalente dentro de la institución. Los pasos anteriores de transformación de información, calibración de un modelo óptimo, predicción para estudiantes en curso y disponibilización de la información han sido automatizados por U-Planner en un producto llamado U-Retention. Este software permite realizar un estudio y seguimiento personal de los estudiantes, donde se realiza una predicción de deserción para cada estudiante de manera individual basada en la evaluación de sus factores de riesgo y que a través de una interface permite disponibilizar esta información de manera sencilla a usuarios como jefes de carrera o servicios estudiantiles, facilitando la tarea de contacto, seguimiento de estudiantes y actualización de la bitácora de contactos. Adicionalmente, U-Retention provee múltiples herramientas de reportería gráfica que permiten comprender el funcionamiento del modelo, analizar el impacto individual de cada factor y estudiar el impacto de la deserción por carrera y cohorte. La aplicación de data science para el estudio de la deserción resulta distinto de su aplicación para la generación de horarios, puesto que en el primer caso la cantidad de información puede ser eventualmente mucho mayor, puesto que debemos considerar la totalidad de los registros históricos, y de mediciones periódicas como asistencia (presencial u online) y evaluaciones.

PRINCIPALES DIFICULTADES

Una primera dificultad en la aplicación de Data Science en la educación es la dificultad de encontrar profesionales del área. En efecto, un data scientist debe manejar simultáneamente conocimientos de programación, bases de datos, visualización y poseer experticia en la industria de aplicación en particular. Dentro del ámbito nacional, no existen carreras donde los estudiantes egresen directamente de data scientists, y sólo muy recientemente han aparecido programas de magíster y diplomados con un sello de data science. Algunas empresas, viendo la dificultad de hallar personas con el exigente perfil anteriormente descrito, han comenzado a formar equipos multidisciplinarios donde cada integrante es experto en una de las áreas anteriormente mencionadas o mantener programas de training destinados a complementar la formación de sus ingenieros. Adicionalmente, los data scientists son altamente requeridos debido a su especificidad y el valor que agregan a sus compañías, por lo que conseguirlos y retenerlos constituye una competición directa con otras compañías, y requiere mantener excelentes condiciones laborales. Una segunda complicación lo constituye la dificultad intrínseca de las distintas tareas habituales de data science. En efecto, las tareas de extracción y conversión de datos requieren una importante cantidad de tiempo dedicado si no se conoce de antemano la mejor estructura de datos para poder dar solución al problema. Adicionalmente, la elección y aplicación de los algoritmos de Machine Learning e Inteligencia Artificial requiere un riguroso análisis, para evitar problemas de sobre-calibración, en los cuales el modelo se ajusta perfectamente a los datos históricos conocidos pero falla al momento de realizar predicciones futuras. Finalmente, las herramientas computacionales en data science y visualización

evolucionan rápidamente y requiere una investigación y experimentación constante para utilizar las mejores librerías computacionales disponibles. Para una Institución de Educación Superior la inversión de tiempo y dinero requerido para obtener un modelo funcional puede superar sus posibilidades, además de alejarlo de sus intereses principales que son la docencia y la investigación. Para llegar a sus modelos actuales U-Planner ha invertido más de 3 años de investigación y desarrollo y llevado a cabo implementaciones en más de 37 Universidades de Latinoamérica. Esto le ha permitido observar, analizar e intervenir distintas realidades, y destilar las mejores prácticas para la gestión del cambio y realizar mejoras en las Universidades mediante la aplicación de data science.

CONCLUSIONES

En este artículo se han discutido algunos aspectos operacionales, como la planificación horaria y asignación de infraestructura, y de aspectos estratégicos como lo es disminuir la deserción de los estudiantes, ambos aspectos que resultan factibles de mejorar mediante la aplicación de técnicas de data science y algorítmica avanzada. Sin embargo, existen desafíos importantes en el aspecto docente cómo lo es la aplicación de data science para apoyar al profesor en el proceso de enseñanza-aprendizaje. Resulta un paso previo el incrementar el volumen de datos recolectados del aula, con importantes requerimientos en el almacenamiento como lo son la seguridad y accesibilidad de la información. Algunas temáticas que ya empiezan a ser exploradas son las rutas de aprendizaje, la evaluación activa de los profesores, el uso de bots para responder preguntas en foros estudiantiles, y la gestión inteligente del aprendizaje de los estudiantes.

La disponibilidad de información ha transformado la relación entre estudiantes y profesores, situación que sólo puede seguir cambiando en los próximos años. En efecto, gran parte de los contenidos que hace algunos años sólo se aprendían en el aula, hoy pueden encontrarse en línea de manera gratuita. Como plantea Jeff Selinger, autor de *College Unbound*, el desafío que tiene las IES es aceptar esta realidad y prepararse para ofrecer una educación del más alto nivel, en la cual el valor agregado sean experiencias educativas irremplazables como el desarrollo de habilidades de trabajo en equipo, el contacto de primera mano con académicos y competencias directamente laborales.

La utilización de data science en educación tiene el potencial de generar grandes cambios dentro de las instituciones, debido a que permite disponibilizar la información de la institución para una mejor comprensión y optimización de sus procesos. Sin embargo, la aplicación de data science es una labor que puede resultar difícil por cuanto requiere conocimientos avanzados y tiempos de desarrollo altos. En efecto, las Instituciones de Educación Superior cuentan con una gran variedad y cantidad de datos, pero que requieren de un proceso de extracción y transformación para un conjunto de datos limpio y consistente que pueda ser utilizado como dato de ingreso en el proceso de modelamiento a algún algoritmo de interés, cuya formulación requiere cuidadoso desarrollo y testeado para lograr predicciones certeras.

REFERENCIAS

Demetriou C. et al. (2011), *Integration, Motivation, Strengths and Optimism: Retention Theories Past, Present and Future*. 7th National Symposium on Student Retention, Charleston, Norman, University of Oklahoma.

Fayyad U. et al. (1996) *From Data Mining to Knowledge Discovery in Databases*. *AI Magazine*, 17(3), 37-54.

Himmel E. (2002), Modelos de Análisis de la Deserción Estudiantil en la Educación Superior. Retención y movilidad estudiantil 91-108.

Lovelace, A. (2010) On the Complexity of Scheduling University Courses, MSc in Computer Science Thesis.

Rodriguez P. et al. (2014), Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile. Concurso BID, nuevos debates: Datos para el desarrollo.

Selingo J. (2013), College (Unbound). New Harvest Publishing

ENLACES

Noticia de Harvard Business Review, Data Science el trabajo más atractivo del siglo 21, <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>.

Noticia de Smart Data Collective, Impacto de big data en la industria minera, <http://www.smartdatacollective.com/rick-delgado/202321/how-big-data-can-impact-mining-industry>

Reporte de Glassdoor, Mejores empleos en America (2016), https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Sitio web U-Planner, <http://www.u-planner.com>