INVESTIGA I+D+i 2016/2017

GUÍA ESPECÍFICA DE TRABAJO SOBRE "Big data"

Texto del Dr. David Rios

Octubre de 2016

1. Resumen

Realizamos una breve descripción sobre los conceptos principales de Big Data y Data Science. Principalmente se han empleado en problemas del sector privado. En el proyecto debéis reflexionar sobre usos sociales de estas disciplinas.

2. Introducción

Las últimas décadas han visto un rápido crecimiento en la capacidad de las empresas para explotar numerosos avances recientes en tecnologías de la información (TI), de la investigación operativa (IO) y la modelización estadística, de cara a recopilar y procesar datos de mercado y de operaciones para apoyar sus procesos de toma de decisiones.

Como resultado, la analítica de negocios (business analytics) se ha convertido en un campo floreciente para la consultoría y la formación empresarial. Sin embargo, mientras que muchas decisiones de algunos gobiernos a menudo vienen apoyadas con métodos tradicionales del análisis de políticas públicas (policy analysis), incluyendo métodos como el análisis de coste-beneficio, pocos departamentos y agencias gubernamentales han logrado, por el momento, aprovechar de forma sistemática grandes masas de datos, evidencia, métodos avanzados de estadística y aprendizaje máquina (machine learning) para informar sus decisiones.

Este hecho constituye una interesante novedad desde una perspectiva histórica, ya que los métodos cuantitativos de ayuda a la toma de decisiones han surgido frecuentemente en el sector público. Por ejemplo, la estadística social, que se remonta a Quetelet, se inició en el siglo XIX para apoyar a los gobiernos, a partir de la idea de que las regularidades estadísticas sugieren señales sobre realidades sociales. Del mismo modo, el campo de la Investigación Operativa nació durante la Segunda Guerra Mundial al servicio de las fuerzas armadas de Estados Unidos y del Reino Unido, y creció rápidamente basado en el desarrollo de métodos de apoyo a la toma de decisiones en problemas militares.

Comparado con la toma de decisiones en el sector privado, los responsables de las decisiones públicas se enfrentan a tareas más complejas. Además de las dificultades propias de los problemas del sector privado, deben enfrentarse a problemas adicionales relacionados con el hecho de tomar decisiones por y para otros. Así, los responsables políticos se enfrentan a las dificultades asociadas a decidir cómo se asignan los recursos públicos, tratándose de decidir "quién obtiene qué, cuándo y cómo". Puesto que tales recursos son escasos, deben tomarse decisiones complejas, como en la famosa disyuntiva “cañones o mantequilla”. Por otra parte, si nos centramos en el caso de sistemas democráticos, algunas de las peculiaridades de la toma de decisiones en el sector público incluirían:

- El público y/o sus representantes toman las decisiones, dependiendo del grado de participación que se contemple;

- Hay funcionarios públicos que, en general, se encargan de la gestión de la organización en la que debe tomarse la decisión;
- El público, en general, es el que paga por el análisis (previo a la toma de decisiones) a través de sus impuestos;
- Sobre el público recaen las consecuencias de la toma de decisiones;
- La valoración última de los resultados de una decisión es típicamente no monetaria; y,
- Los métodos utilizados para informar sobre las decisiones están sujetos al escrutinio público.

Otras cuestiones que diferencian las decisiones públicas frente a las privadas se refieren a la coexistencia de distintos sistemas de valores y culturas, y que tales decisiones pueden verse afectadas por los cortos horizontes electorales de los representantes, con el consiguiente riesgo asociado a una visión cortoplacista de lo público. Además, puede haber estructuras organizativas más burocráticas que dificultan la toma de decisiones.

3. Analítica y analítica de negocios

Ya hemos mencionado los orígenes de la Estadística y la Investigación Operativa en la política pública. Durante la última década, el crecimiento de la potencia de cálculo y los avances en tecnologías de grandes datos (Big Data) han proporcionado nuevas perspectivas en tales disciplinas, lo que ha llevado a una nueva visión denominada Analítica (Analytics), que está demostrando ser extremadamente valiosa para ayudar a los responsables de toma de decisiones en áreas de negocios y la industria.

La Analítica puede centrarse en enfoques descriptivos, predictivos o prescriptivos. Típicamente, apoya el descubrimiento y la presentación de patrones significativos en grandes conjuntos de datos, en problemas con gran cantidad de información registrada, para cuantificar, describir, predecir y mejorar los resultados de una organización. Cuando nos referimos al entorno de negocios, se denomina analítica de negocios (business analytics), un término popularizado en los últimos años. A menudo, combina métodos de la estadística, la investigación operativa, el aprendizaje de máquina y la informática, junto con disciplinas como la sociología, la psicología y la economía. Las ideas proporcionadas por los datos se emplean para recomendar acciones y guiar la toma de decisiones y la planificación en organizaciones. Sus resultados pueden emplearse como entrada a la toma de decisiones por personas; también pueden alimentar sistemas automáticos de ayuda a la toma de decisiones. Por contraste, el ya más tradicional concepto de inteligencia de negocio (business intelligence) tiende a referirse a la extracción de información, la elaboración de informes, y la provisión de alertas en conexión con el problema aplicado de interés.

En la industria, el énfasis en el área de la Analítica se ha puesto en tratar de resolver problemas relacionados con el análisis de conjuntos de datos masivos y complejos, frecuentemente, en entornos muy cambiantes, más allá de la evolución y el desarrollo de ERPs y almacenes convencionales. Tales conjuntos de datos suelen denominarse Big Data y se caracterizan por tres rasgos típicos de los negocios que emplean sistemas

de transacciones online y, en consecuencia, permiten amasar grandes volúmenes de datos rápidamente:

- Grandes volúmenes de datos generados. Como ejemplos, Walmart acumula más de 2,5 petabytes por hora de transacciones de clientes; Facebook recoge 300 millones de fotos y 2,7 millones de Likes por día. Aproximadamente, 5 exabytes al día se almacenan en la actualidad.
- Gran heterogeneidad de los datos generados, que pueden provenir de fuentes tales como mensajes de blogs, imágenes en redes sociales, correos electrónicos, archivos PDF, datos geoespaciales, lecturas de sensores en una ciudad, o señales GPS en teléfonos móviles. De hecho, podríamos contemplarnos a cada uno de nosotros, hoy en día, como generadores permanentes de datos, a partir de nuestras interacciones con smartphones.
- Datos generados a gran velocidad. En muchos casos, la velocidad de generación tiende a ser más importante que el volumen de datos generados, en el sentido de que se deben tomar decisiones en tiempo real, teniendo que evaluarse la información en tiempo real, a partir de datos obtenidos también en tiempo real.

El análisis de estos tipos de datos no estructurados (no muestreados) constituye un reto importante en la industria, dando lugar a nuevos paradigmas como la Ciencia de Datos y la Ingeniería de Datos. Los datos no estructurados difieren de los que lo son en que su formato es muy variable y no pueden ser almacenados en bases de datos relacionales tradicionales sin esfuerzos significativos que impliquen transformaciones de datos complejas. Se emplean así bases de datos No SQL más escalables, con ejemplos como CouchDB, MongoDB, Neo4J y Riak. Manejar tales masas de datos requiere marcos que permitan realizar cálculos sobre grandes cantidades de datos, como MapReduce y su implementación Hadoop (debida a Google), que facilita el procesamiento distribuido sobre conjuntos más pequeños de datos. Finalmente, necesitamos también infraestructuras de almacenamiento sobre Hadoop que faciliten el resumen de datos y sus análisis como Hive (debido a Facebook). Dentro de estos desarrollos tecnológicos, deberíamos mencionar Python, como el lenguaje principal de programación con propósito numérico, así como R, para inferencia y predicción.

Además de los avances tecnológicos, también hay nuevas clases de métodos de análisis de datos que permiten la extracción de información de conjuntos masivos de datos. Estos van más allá de técnicas tradicionales, como los modelos de regresión, los modelos de series temporales, los clasificadores de k-vecinos más cercanos, llegando a métodos más recientes como los árboles de clasificación y de regresión, los conjuntos de máquinas o las máquinas de soporte vectorial (support vector machines). Con frecuencia, éstos requieren nuevas implementaciones como en el caso de las funciones en R biglm y bigmemory para la regresión lineal en lugar del tradicional lm. Otro ejemplo es Mahout, que facilita la clasificación y el análisis de conglomerados sobre Hadoop. El análisis de redes sociales, con orígenes en la sociología, y otros métodos analíticos para estructuración y extracción de significado, como los mapas cognitivos también están ganando importancia, debido a los datos provenientes de redes sociales.

En cualquier caso, los datos parecen ahora más accesibles a los gestores, que tienen una gran oportunidad para tomar mejores decisiones utilizándolos para aumentar ingresos, reducir costes, mejorar el diseño de productos, detectar y prevenir el fraude, o la mejora de la participación de clientes a través del marketing personalizado. Esto ha conducido a un nuevo concepto de empresa que toma decisiones basadas en la evidencia, con ejemplos claros como Google, Facebook, Amazon, Walmart y algunas de las líneas aéreas más avanzadas.

4. Analítica para políticas públicas

Hemos descrito cómo el crecimiento de los datos ha conducido a nuevos desarrollos tecnológicos y científicos en lo que ahora se llama Analítica. Cuando se aplica a los negocios se denomina Analítica de Negocios. De hecho, este término se ha vuelto tan popular que hay numerosas universidades que ofrecen ya estudios sobre estos temas.

Las mismas cuestiones en relación con la disponibilidad y el posible uso de datos se están encontrando en el contexto de las políticas públicas, como, por ejemplo, en el caso de ingresos hospitalarios, registros médicos electrónicos, datos meteorológicos, venta de propiedades, registro de votantes, datos procedentes de cámaras de vigilancia o posicionamiento de teléfonos móviles, que pueden ser fuentes tremendamente útiles de grandes cantidades de datos para numerosos departamentos gubernamentales. Además, estas fuentes coexisten con las más tradicionales procedentes de sistemas pre-diseñados de recopilación masiva de datos, que incluyen los censos, los documentos de registro de recaudación de impuestos o, finalmente, los distintos sondeos que realizan los gobiernos.

Así, podríamos pensar en aplicar la Analítica para apoyar la toma de decisiones en la elaboración de políticas públicas, lo que lleva, de forma natural, al concepto de Analítica para Políticas (Policy Analytics). Este es un nuevo término acuñado en la literatura científica en trabajos de Tsoukias y colaboradores. Sin embargo, cuando se realiza una búsqueda de ese término, vemos que aparecen sólo un par de empresas con nombre relacionado (Policy Analytics, Public Policy Analytics) y empresas como Oracle, Booz-Allen-Hamilton o IBM ya han incluido el término dentro de su cartera de actividades. Carnegie Mellon tiene también una sección de Analytics políticas dentro de su programa de Políticas Públicas. Pero, como mencionamos en la introducción, pocas decisiones gubernamentales se benefician aún del uso sistemático de grandes masas de datos y técnicas avanzadas de modelización.

Por comparación con las aplicaciones industriales, no es difícil vislumbrar las enormes aplicaciones potenciales que tienen en problemas como examinar la distribución de patrones de sucesos de salud, el desarrollo racional de planes de infraestructura, el empleo del conocimiento sobre comportamiento para promover la eficiencia energética, el desarrollo de servicios personalizados de gobierno, la mejora de la experiencia en visitas turísticas, la identificación de barrios con servicios sociales inadecuados, el diseño de ciudades inteligentes, entre otros muchos.

Con frecuencia, se tendrá que utilizar una gran variedad de analíticas en todo el ciclo del análisis de políticas (policy analysis cycle) Así, por ejemplo, en la fase de establecimiento de la agenda, se puede emplear la minería de textos o el mapeo coginitivo (cognitive mapping); modelos de optimización, simulación, y decisiones multicriterio para la fase de análisis; la planificación participativa en la fase de toma de decisiones; los modelos de asignación de recursos y de optimización de operaciones en tiempo real en la fase de implementación; y, finalmente, distintos métodos de evaluación como sistemas de medición inteligente o SIGs participativos en la fase de monitorización. También pueden adaptarse a una variedad de ideologías y procesos de formulación de políticas, incluyendo grados variados de participación-representación.

5. Cuestiones de debate

Las siguientes cuestiones pueden ayudarte a organizar tu investigación:

- ¿Por qué han empezado a ser relevantes estas disciplinas en tiempos recientes?
- ¿Cuáles son las principales tecnologías y metodologías empleadas en Ciencia e Ingeniería de Datos?
- ¿Por qué hay un predominio, por el momento, de estas disciplinas en el sector privado?
- ¿Qué fuentes de datos (de origen público y privado) son relevantes en la toma de decisiones públicas?
- ¿En qué campos de aplicación en política pública podrías aplicar Ciencia de Datos?
- ¿Qué aspectos éticos y de privacidad deberías tener en cuenta en un proyecto Big Data aplicado al sector público?
- ¿Cuál sería un ciclo de análisis de políticas públicas y cómo pueden ayudar los métodos de ciencia de datos.
- Identifica un problema social de interés y describe cómo aplicar el ciclo anterior.
- Diseña un modelo de negocio alrededor del problema que has descrito.

6. Fuentes de información

La mayoría de las fuentes de información relevantes se encuentran en inglés.

Aquí tenéis algo de información de wikipedia

- https://es.wikipedia.org/wiki/Ciencia_de_datos
- https://es.wikipedia.org/wiki/Big_data
- https://en.wikipedia.org/wiki/Policy_analysis (no hay versión en español de ésta)

Información de IBM

- https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Tenéis un glosario en

- http://fundacionbigdata.org/glosario-big-data/

El blog de Soraya Paniagua es interesante

- http://www.sorayapaniagua.com/2011/11/01/la-ciencia-de-los-datos-bdii/

Una breve introducción a la presencia de Big Data en el INE

- http://www.ine.es/ss/Satellite?L=es_ES&c=INEmasNoticia_C&cid=1259948778832&idp=1254736092060&pagename=masINE%2FmasLayout

Ejemplos de uso en BBVA

- http://www.centrodeinnovacionbbva.com/proyectos/big-data