# Big data: la fiebre del siglo XXI

## Reinventándose la estadística para los nuevos datos

istock

"Las Matemáticas y la Informática se han impuesto en asuntos humanos de vital importancia; solo hay que saber los sesgos que el procedimiento puede conllevar. Ése es nuestro reto".

Pilar Olave Rubio

www.scitecheuropa.eu

Cuando las personas realizamos actividades diarias a través de internet, como navegar por la web o compartir contenido en redes sociales, dejamos un rastro de datos y, en los últimos años, en cualquier entorno se han generado tales cantidades de datos que cualquier proceso puede ser estudiado y optimizado con algoritmos analíticos. Algunas preguntas, que es interesante responder en este contexto, serían: ¿son eficientes los algoritmos para todos los tipos de datos y grupos sociales?, ¿cómo se codifican "situaciones", "sentimientos"? y ¿cómo afectan a los algoritmos?, ¿sabemos generar *valor* en dichos datos?, pero *valor* para todos esos grupos, no solo para los que gestionan y generan los algoritmos. Hay que tener presente que la automatización de los procesos impacta en cantidad de procedimientos (selección de personal, casa, seguros, evaluaciones en el trabajo,….) y ello puede aumentar las desigualdades sociales aún más si cabe. La cuestión previa que debemos tener presente es que los datos son ahora de tamaño muestral ingente, muy complejos, grabados en diferentes soportes y de alta dimensión; es lo que se conoce habitualmente como Big Data y los tratamientos estadísticos deben adecuarse a esta nueva estructura de datos. Y, lo que es también esencial, decidir cuáles son los datos que utilizaremos y cuáles no, atendiendo no solo a principios estadísticos, sino morales. En la era del Big Data, un programa de ordenador es capaz de procesar miles de currículos, préstamos,...y clasificarlos. Las Matemáticas y la Informática se han impuesto en asuntos humanos de vital importancia; solo hay que saber los sesgos que el procedimiento puede conllevar. Ése es nuestro reto. En este artículo reflexionaré sobre algunos de estos puntos, y trataré de plantear los cambios que se avecinan en los procedimientos estadísticos más usuales.

"No es la especia más fuerte la que sobrevive, ni la más inteligente, sino la que mejor responde al cambio."

Charles Darwin

Las grandes compañías tales como Facebook, Netflix, Amazon,… combinan análisis de datos masivos con aprendizaje automático para ofrecer experiencias personalizadas, gracias a un profundo conocimiento del *dato disponible.* Todos hemos visto aparecer en nuestras pantallas la serie o película que Netflix, HBO,… nos sugiere según el rastro de nuestras anteriores visitas. Pero, como no se puede gestionar lo que no se mide, será de obligado cumplimiento para esas empresas saber lo que se busca en web, quien lo observa, lo que influyen promociones de ventas, los diseños de páginas web, etc. Gracias a los teléfonos móviles, redes, ubicaciones, fotos, se producen torrentes de datos difíciles de manejar (con mucho ruido pero también con determinadas señales que generan *valor*) y, ya que la mayoría de datos disponibles son desestructurados (no están organizados en una sola base de datos), nos obliga a *desarrollar métodos y algoritmos para realizar proyecciones futuras.* En Internet se generan enormes cantidades de datos en términos de búsqueda de entradas, foros de Internet, registros de chat y mensajes de microblogs. Esos datos están estrechamente relacionados con la vida diaria de las personas y tienen características similares; pueden carecer de valor individualmente pero, a través de la explotación acumulada de Big Data, se puede identificar información útil como hábitos y aficiones de los usuarios, e incluso es posible pronosticar las conductas y los estados de ánimo de ellos. Se estima que el volumen de datos comerciales de todas las empresas del mundo puede duplicarse cada año. Uno de los valores más importantes de una empresa es su llamado capital intelectual. Este se define como el conjunto de aportaciones no materiales, que en la era de la información se entienden como el principal activo de las empresas del tercer milenio. Una forma de explotar el capital intelectual que está teniendo más auge en los últimos años es el análisis de contenidos, herramienta de investigación utilizada para determinar la presencia de palabras clave o conceptos dentro de un texto o grupo de textos y las empresas cada día lo explotan más para saber qué palabras /conceptos / ubicaciones en red les interesa más.

Por citar un ejemplo muy sencillo, a los solicitantes de un préstamo y/o trabajo se les puede pedir que describan motivos de dicha petición, pues no solo son predictores de la probabilidad de pago/cumplimiento en el trabajo las variables usuales (CV, situación económica,..), sino que el lenguaje presencial o en redes sociales es una variable de interés en un análisis financiero, laboral, etc… Así pues, en un futuro muy próximo, un usuario

que busca un préstamo y/o trabajo debería preocuparse no solo por su historial académico y/o financiero sino además por su actividad on line. No obstante, hay que tener en cuenta que las variables utilizadas por los algoritmos son valores sustitutivos o proxies y muchas veces injustos. Los demandantes de empleo/préstamo deben organizar sus cv pensando en los lectores automáticos y salpicar sus currículos de palabras relacionadas con el objetivo de su demanda (para más detalles, consultar la página: *mashable.com/*2012/05/27. "*12 ways to optimize your resume for applicant tracking systems*"). Peter Norvig, director de investigación de Google en 2009, lo expresa insistentemente: "No tenemos mejores algoritmos. Solo tenemos más datos". Gary King, director del Instituto de Ciencias Sociales y Cuantitativas de Harvard, ha declarado recientemente que la toma de decisiones basada en enormes fuentes de datos se extenderá a todos los ámbitos: académico, negocios, medicina, administraciones, etc…

Los ***nuevos datos*** son ahora el oro del siglo XXI. Pero realmente, ¿mejoran los negocios, nuestra vida…? Veámoslo en un sencillo ejemplo: Una importante aerolínea americana supo que aproximadamente el 10% de los vuelos a un importante aeropuerto tenía, al menos, una brecha de 10  minutos entre *tiempo estimado* de llegada y *tiempo real* de llegada del vuelo, y un 30% de vuelos, al menos un brecha de 5 minutos. Para buscar soluciones, la compañía recurrió en 2001 a PASSUR Aerospace, expertos en tecnologías de Big Data en aviación, y recalculó los tiempos estimados combinando los datos disponibles de aviación con otros de meteorología, es decir con otros factores que afectan a las estimaciones, siendo sus fuentes una extensa red de estaciones de radar pasiva, instaladas cerca de los aeropuertos de interés. Esto produjo una ***entrada masiva de datos*** en la red, y lo que lo convierte en un problema Big Data es que cuenta con esa información multidimensional a lo largo del tiempo. En 2012, PASSUR tenía


istock


pixabay.com

más de 155 instalaciones por lo que produce una inundación de datos digitales en tiempo y espacio. A partir de la información obtenida se crearon *patrones de aterrizaje* en función de las condiciones meteorológicas, y así se redujeron las brechas entre tiempos estimados y reales. Para más detalles, McAfee A. and Brynjolfsson E. (2012) "*Big Data: The management revolution*".

Otro ejemplo que subraya el interés de resolver un problema a través de técnicas de Big Data es el que tuvo lugar durante la pandemia de gripe de 2009. Google obtuvo información valiosa de fácil acceso y de mayor trascendencia que la proporcionada por los pacientes, ya que se descubrió que, durante la propagación del virus, *las entradas buscadas por muchos enfermos* eran diferentes a las habituales, y las frecuencias de uso de dichas entradas se correlacionaron con la propagación del virus N1H1. En un principio esta estrategia falló pues la psicosis suele provocar que todo el mundo consulte en red y no se sepa realmente donde están las consultas de focos reales. Posteriormente el sistema se sofisticó y la plataforma de gestión de datos masivos InfluenzaNet  resultó segura, flexible y rápida. Así se encontraron grupos que *eran muy relevantes para detectar cómo se propagaba el brote de N1H1*, tanto en tiempo como en ubicación, a través de implementar modelos matemáticos. No tenemos algoritmos perfectos,

"Los nuevos datos son ahora el oro del siglo XXI. Pero realmente, ¿mejoran los negocios, nuestra vida…?."



**Cambridge Analytica usó datos de Facebook para influir a los votantes a favor del Brexit.**

> “Google tenía mucha más información que la proporcionada por ningún sondeo, a nadie se le hubiera ocurrido pensar años antes en Donald Trump como un candidato presidencial.”

pero la próxima vez que ocurra una nueva pandemia, el mundo tendrá una mejor herramienta a su disposición para predecir y por lo tanto prevenir su diseminación.

En los albores del siglo XXI se acuña el término Big Data, que hace referencia al estudio de enormes volúmenes de datos complejos/desestructurados con múltiples fuentes autónomas y en diferentes soportes. Es decir, no estamos hablando solo de flujos de datos sino de ***nuevos datos*** que no pueden ser almacenados y/o procesados con un ordenador de capacidad tradicional. En este contexto el planteamiento analítico difiere al de la *inferencia estadística tradicional*, pues los datos ahora no provienen de una muestra convenientemente seleccionada para el objetivo de nuestra investigación, sino que los datos se autoseleccionan, por lo que nos podemos encontrar con graves problemas de sesgos ocultos. Y, aunque el objetivo final sea el mismo que hace 30 años, es decir, *obtener características y/o patrones para las poblaciones de interés,* ***la metodología debe adaptarse a esta nueva situación de información masiva***.

Este articulo *pretende destacar el interés de las técnicas de Big Data*, es decir, de aprendizaje automático a través de los datos , usando metodología estadística, así como mostrar algunos inconvenientes de los procedimientos automatizados y las posibles desigualdades que en determinados entornos puede conllevar, ya que el *conocimiento generado* se utilizará de forma estratégica para formar juicios, valores y tomar decisiones que afectarán de forma desigual a diferentes grupos, y eso puede generar ***sesgos*** que, si no se detectan convenientemente, pueden depreciar el valor de dichas técnicas, entendido el término *valor* como principio moral. Por citar un ejemplo cercano, el escándalo de Cambridge Analytica (CA) dejó claro que la publicidad y comentarios en determinados medios de comunicación, así como presiones sociales en la época previa al referéndum del Brexit, fueron muy diferentes para los perfiles creados por la consultora CA. Es decir, gracias a la utilización de técnicas estadísticas para datos masivos digitalizando previamente comentarios, fotos y “me gusta” de los usuarios de Facebook (en distintas áreas de Reino Unido), se crearon los perfiles citados. Posteriormente, los impulsores de “la salida” plantearon estrategias de marketing y publicidad muy diferentes para los grupos establecidos. El escándalo sobre esa utilización de datos personales, así como las grandes sumas que se invirtieron, es por todos conocido. De la misma forma, en las primarias republicanas de 2016, los analistas de sondeos concluyeron que Trump no tenía ninguna posibilidad de ganar. Pero, los indicios de que podía ganar estaban en internet, Google tenía mucha más información que la proporcionada por ningún sondeo, a nadie se le hubiera ocurrido pensar años antes en Donald Trump como un candidato presidencial, pero en las búsquedas en red se pudo ver cómo sus ataques a la inmigración fue una de sus principales bazas. Las búsquedas en Google demostraron que el racismo persiste en gran parte de americanos y se pudo establecer un mapa de su ubicación que sería vital para explicar el éxito de Trump. Así pues, nos podemos preguntar: las búsquedas en Google, ¿son óptimos predictores? Bueno, aún queda mucho trabajo estadístico por delante, pero desde luego tiene un potencial, yo diría, diferente que

el de los sondeos clásicos. El aprendizaje automático de los datos será óptimo para la sociedad civil en su conjunto cuando los sesgos en los datos sean fáciles de detectar o de incorporar a los modelos. Desde mi punto de vista, este es uno de nuestros retos.

Ahora vamos explicar algunos cambios cualitativos producidos en los métodos de *análisis de datos:* ***nuevos métodos para nuevos datos.***

La gran pregunta es: ¿qué hay que transformar en el binomio Estadística /Datos? Vamos a explicarlo brevemente y lo haremos en tres escenarios: ***Acceso a los datos, conocimiento de su dominio y realización del método y algoritmo adecuado***. El desafío en el ***acceso a la información*** es desarrollar los procedimientos informáticos para obtener datos de diferentes ubicaciones, ya que ahora muchos de los datos son generados de forma automática y pueden ser: textos, imágenes, videos, audios, con diferente frecuencia y periodicidad. Una representación inadecuada de los datos reducirá el valor de los datos originales e incluso puede obstruir el análisis efectivo de los datos. Una representación de datos eficiente debe reflejar la estructura, clase y tipo de datos, así como las tecnologías integradas, para permitir operaciones eficientes en conjuntos de datos de diferentes fuentes. Hay unos avances relativamente lentos en los sistemas de almacenamiento, que no pueden soportar datos tan masivos. Por lo tanto, hay que *decidir qué datos se almacenarán y cuáles se descartarán. Además no hay que olvidar que los nuevos conjuntos de datos nos ofrecen un número de variables mayor que los datos tradicionales y eso aún complica más la selección de datos.*

Posteriormente, se desarrolla el software necesario para intercambios de información entre productores de datos, así como resolver los ***problemas*** de privacidad y/o semántica ***de muchos dominios***, ya que son las personas que han planteado la investigación las que conocen profundamente ese dominio (el área funcional de la empresa que lo va a adoptar, sector de actividad económica, etc.), y no solo eso sino que los usuarios, al no poseer en muchas ocasiones los datos, tienen que contratar a auditores de datos para, a través de mecanismos de cifrado, protegerlos. La mayoría de los proveedores de servicios de Big Data en la actualidad no podrían mantener y analizar de forma efectiva los conjuntos de datos tan grandes debido a su capacidad limitada, por ello deben confiar en profesionales para analizarlos, lo que aumenta los riesgos potenciales de seguridad. Así, se debe establecer una estructura interna para ayudar a los expertos en diversos campos (métodos cuantitativos y computación) a utilizar plenamente su experiencia, a fin de cooperar para completar los objetivos analíticos.

> “Nuevas técnicas estadísticas para datos masivos pues hay que tener presente que datos masivos no implica muchas veces más información.”

**Finalmente, se desarrollaran las *técnicas estadísticas adaptadas con fiabilidad* a ese contexto** y es aquí donde el aprendizaje automático de los datos buscando los modelos y / o algoritmos, lo que en estos últimos años se ha convertido en un objetivo de los ***nuevos métodos estadísticos.*** El análisis de Big Data involucra principalmente métodos analíticos para datos tradicionales y arquitectura analítica para Big Data y software utilizado para minería y análisis de Big Data. *Estamos hablando de hacer la ciencia estadística más convergente con las ciencias de la computación.* Para más detalles, Galeano P. and Peña D. (2019) “Data science, Big Data and Statistics”.

Muchos métodos de análisis de datos tradicionales todavía pueden ser utilizados para el análisis de grandes masas de datos. Podemos destacar, entre otros: Las técnicas de *análisis exploratorio de datos* que han tomado ahora una presencia importante en la metodología estadística frente a la inferencia de datos. Pero, al explorar los datos masivos nos encontramos con mucha frecuencia que una representación gráfica muy útil, como es el diagrama de dispersión, se convierte en una búsqueda compleja al no


www.iebschool.com

poder distinguir patrones subyacentes. En otro contexto, en el área biomédica un individuo se puede representar utilizando información demográfica simple: sexo, edad, historia familiar,...o añadir información visual tales como imágenes y secuencias de expresiones de microarrays. Estas son algunas de las características nuevas a las que los analistas nos enfrentamos. *Análisis de correlación:* es un método analítico para determinar las relaciones entre los fenómenos observados y, en consecuencia, la realización de pronósticos y controles, pero al explorar sin modelo subyacente, las relaciones no son causales y de difícil interpretación. *Análisis clúster:* método estadístico para agrupar y clasificar objetos de acuerdo con algunas características, buscando una alta homogeneidad dentro de cada grupo, pero en el contexto de datos masivos el análisis de conglomerados clásico tiene muchas limitaciones pues la creación de grupos con objetos similares en él y heterogéneos respecto a los demás es ahora muy complejo, ya que la aparición de elementos atípicos se convierte en algo habitual. Hay que tener presente que la homogeneidad no se mantiene en el tiempo y los datos masivos se obtienen longitudinalmente. *Tiene más sentido buscar direcciones de proyección que muestren de forma clara la heterogeneidad de la muestra y posteriormente buscaremos los grupos o clúster sobre las proyecciones. Análisis factorial*, procedimiento de reducción de dimensión clásico que ahora se convierte en imprescindible. La maldición de la dimensionalidad es ahora un elemento omnipresente en este tipo de datos. *Algoritmos de minería* de datos, proceso para extraer información y conocimiento ocultos, desconocidos, pero potencialmente útiles, de datos masivos, incompletos, ruidosos, difusos y aleatorios, son ahora de vital importancia.

Uno de los problemas que se pueden manifestar con frecuencia en datos masivos son las conclusiones extraídas en ***poblaciones heterogéneas***, pues en Big Data este no es un caso particular de los datos sino que es una cuestión estándar. Es decir, nos encontraremos con la *famosa* ***paradoja de Simpson***, en que las variables escondidas nos llevaran a inferencias incorrectas y, si en muestras pequeñas su efecto es importante, en muestras grandes dicho efecto se manifiesta de forma ***casi automática***. Por citar un ejemplo elemental (Peña D. *“Big Data y Estadística”* Workshop on Big Data and Statistics, Universidad Carlos III, 2015), si de 4000 solicitudes (2000 hombres y 2000 mujeres) para ingresar en una universidad, los admitidos han sido 1140 (57%) en hombres y 960 (48%) en mujeres, concluiremos que hay sesgos de admisión por sexo.

Pero si desglosamos por solicitudes en las correspondientes Facultades (Humanidades, Ingeniería y Economía), es fácil encontrar ***preferencias por sexo*** en cada

| Universidad | Admitidos | No admitidos | Admisión |
|---|---|---|---|
| Hombres | 1140 | 960 | 57% |
| Mujeres | 960 | 1140 | 48% |

| Humanidad | Admitidos | No admitidos | Admisión |
|---|---|---|---|
| Hombres | 225 | 75 | 75% |
| Mujeres | 560 | 240 | 70% |

| Ingenierías | Admitidos | No admitidos | Admisión |
|---|---|---|---|
| Hombres | 140 | 560 | 20% |
| Mujeres | 36 | 164 | 18% |

| Economía | Admitidos | No admitidos | Admisión |
|---|---|---|---|
| Hombres | 590 | 410 | 59% |
| Mujeres | 540 | 460 | 54% |

especialidad y no hay ningún sesgo en la admisión. Este ejemplo refleja que las variables escondidas nos las vamos a encontrar en datos masivos en donde la distribución de la población o poblaciones subyacentes no se conoce y es muy fácil caer en este tipo de errores.

En datos masivos hay que ser muy cuidadosos por desconocimiento del investigador de las distribuciones condicionadas, según variables.

Estamos pues, ante lo que ya he citado: ***nuevas técnicas estadísticas para datos masivos*** pues hay que tener presente que datos masivos no implica muchas veces más información.

Otro problema clave, en el universo de datos y en redes sociales, es detectar elementos virales así como ***elementos atípicos*** pues en datos masivos esos elementos son referentes de posibles existencias de *otros modelos*. El sesgo en los datos se presenta como un problema considerable y es frecuente en un contexto de Big Data. Los *datos se autoseleccionan, por lo que no existe un procedimiento de muestreo controlado por el investigador,* que permita garantizar la representatividad de ese gran volumen. En este caso se puede fragmentar el problema, es decir, convertir un análisis Big Data en muchos análisis de muchas muestras más pequeñas. Precisamente los métodos de submuestreo se presentan como una herramienta de visualización y análisis en los casos en que el tamaño muestral haga impracticable gráficos sencillos, cuestión que ya he mencionado anteriormente.

Así pues, a las personas que nos dedicamos al apasionante mundo de la Ciencia Estadística se nos plantean cambios de paradigma en lo que respecta a afrontar nuevos métodos/retos de la mano de la computación con nuevas bases de datos y técnicas computacionales de altas prestaciones. Debemos tener claro nuestro papel en esta Ciencia de los Datos y es lo que quiero plantear para finalizar estos breves trazos. El objetivo del Big Data es encontrar patrones invisibles a primera vista. Este es el reto de los científicos de datos, sugerir candidatos óptimos para un trabajo, un préstamo, riesgos de sufrir una enfermedad, etc. Pero se nos avecina un grave problema ético, por muy alta que sea la capacidad predictiva de los modelos estadísticos, ¿es adecuado utilizar análisis de contenidos para evaluar a un candidato para un determinado puesto según cuándo y dónde ha buscado determinados hechos, cosas, chistes, vocablos,...? Estamos ante una gran revolución basada en datos, pero eso no quiere decir que cualquier pregunta se puede contestar siendo el input *exclusivamente datos.* En primer lugar porque el científico de datos ha tenido previamente que codificar la información y además esos modelos no eliminan la necesidad de utilizar las demás formas de entender el mundo que hemos desarrollado a lo largo de milenios. Estos modelos no deben construirse solo con datos (incluso pensando en una óptima codificación de las acciones humanas) sino, como he sugerido en el principio de estas reflexiones, desde un vasto conocimiento estadístico, sabiendo qué datos son los más adecuados en cada contexto.

A mis alumnos de Marketing e Investigación de Mercados.

Pilar Olave Rubio
Métodos Cuantitativos para Economía y Empresa
Facultad de Economía y Empresa
Universidad de Zaragoza


**REFERENCIAS**


- P. Galeano and D. Peña (2019) "Data Science, Big Data and Statistics" TEST, 28: 289-329.
- A. Halevy, P. Norvig and F. Pereira (2009) "*The Unreasonable Effectiveness of Data*" IEEE Intelligent Systems 24 (2): 8-12
- (2012/05/27). "*12 ways to optimize your resume for applicant tracking systems*". https://mashable.com
- A. McAfee and E. Brynjolfsson (2012) "*Big Data: The management revolution*" Harvard Business Review
- C. O'Neil (2018) "Armas de destrucción matemática". Capitán Swing
- D. Peña (2015) "*Big Data y Estadística*" Workshop on Big Data and Statistics, Universidad Carlos III
- S. Stephens-Davidowitz (2019) "Todo el mundo miente. Lo que Internet y el *Big Data* pueden decirnos sobre nosotros mismos". Capitán Swing