

Desmitificando los algoritmos.preciando los datos, el pasado, los valores sustitativos

Resumen:

El presente artículo, aunque mencione los inconvenientes de los algoritmos, o los peligros de las redes sociales, pretende destacar algunos de los aspectos de los algoritmos con un enfoque desde el análisis de inteligencia, relacionado con la toma de decisiones, detallando cómo operan y en qué se basan.

De esos aspectos se desprende la importancia de su materia prima: los datos, y cómo infinidad de veces se recurre a los que hay disponibles, los valores sustitativos, que muchas veces no son los óptimos para representar el concepto estudiado y, por tanto, pueden distorsionar el resultado final.

A pesar de ello, referencias, valores y secuencias similares a los valores sustitativos son utilizados en otros enfoques de carácter probabilístico, también para conocer el presente y predecir el futuro, como es el caso de los superpronosticadores.

En ambas ocasiones se recurre a la búsqueda de cantidades y cifras, en definitiva, a números y su gestión.

Palabras clave:

Algoritmo, datos, valor sustitutivo, superpronosticador.

***NOTA:** Las ideas contenidas en los *Documentos de Opinión* son responsabilidad de sus autores, sin que reflejen necesariamente el pensamiento del IEEE o del Ministerio de Defensa.

Demystifying the algorithms. Appreciating data, past and proxies

Abstract:

This article, although it mentions the disadvantages of algorithms, or the dangers of social networks, aims to highlight some of the aspects of algorithms with an approach from intelligence analysis, related to decision making, detailing how they operate and in what are they based on.

From these aspects, the importance of its raw material can be deduced: the data and how many times we resort to those that are available: the proxies, which are often not optimal to represent the concept studied and therefore can distort the result.

Despite this, references, values, and sequences like the proxies are used in other probabilistic approaches, also to know the present and predict the future, as is the case of the super forecasters.

On both occasions, the search for quantities and figures is used, in short, numbers and their management.

Keywords:

Algorithm, data, proxies, superforecasting.

Los datos, recursos enormemente abundantes, fáciles de obtener, pero extremadamente valiosos

«Todo lo que no se mide, no me interesa»
Iván Redondo¹

Los metales preciosos, los recursos energéticos, el agua, los alimentos, los valles más fértiles, los puertos naturales más resguardados siempre han sido objetos en disputa. En las últimas décadas, nuevos elementos han engrosado esta lista: el coltán, las tierras raras, las nuevas rutas marítimas, el espacio..., aunque un elemento inmaterial, «revolucionario» y ubicado en el ciberespacio, despunta con entidad propia: los datos, millones de millones de datos, extremadamente fáciles de obtener, que pueden ser procesados gracias a los avances informáticos y que convierten en supermillonarios y superpoderosos a quienes los recopilan y gestionan.

Todos somos conscientes del inmenso poder de las multinacionales tecnológicas y en general de la tecnología, sin embargo, muchos expertos afirman que los datos son el recurso más valioso hoy en día, más que el oro y el petróleo. Si admitimos que los datos son uno de los elementos estratégicos más valiosos, deberíamos priorizar su análisis también en los centros de estudios estratégicos, trascendiendo los ámbitos especializados en *big data*.

Los datos se generan al registrarnos en un sitio web, o en cualquier interacción que efectuemos en Internet, ya sea en un buscador, con nuestra tarjeta de crédito o de puntos de un supermercado, redes sociales propias (e incluso de nuestros contactos), etc. Todo queda registrado con un dato, difícil o imposible de borrar, que se asocia a nosotros. Los datos son los componentes esenciales de los algoritmos. Muchas empresas venden los datos de sus usuarios, existiendo compañías que se dedican a la compraventa de bancos de datos. Así, en noviembre de 2018 la ONG alemana Tactical Tech ofrecía una lista de 329 compañías especializadas en campañas electorales².

¹ VALLS, Fernando. «Un día de trabajo con Iván Redondo: así funciona la sala de máquinas de Moncloa», *lainformación.com*, 14 de julio de 2019. Disponible en: <https://www.lainformacion.com/espana/ivan-redondo-sanchez-moncloa-psoe/6506465/> [Todos los enlaces de este artículo han sido comprobados el 1 de marzo de 2021].

² «Who's Working for Your Vote?», *Tactical Tech*, 29 de noviembre de 2018. Disponible en: <https://ourdataourselves.tacticaltech.org/posts/whos-working-for-vote/>

Refiriéndose a las mejores mentes de una generación, la gran experta Marta Peirano afirmaba: «Los que no trabajaban en Facebook, Google, Twitter, LinkedIn, Amazon o Groupon buscando maneras de mejorar los anuncios estaban en Wall Street escribiendo algoritmos capaces de digerir cantidades industriales de datos de mercado para tomar decisiones de compraventa en microsegundos»³.

Los datos despiertan el interés, no solo por su explotación comercial en sectores relacionados con la publicidad, el marketing u operaciones de alta frecuencia en bolsa; hace tiempo dieron el salto a otros sectores aparentemente alejados, como puedan ser las ciencias sociales o la política (como el listado de Tactical Tech confirma) y esto ha sido así porque habían demostrado su utilidad en esos otros campos. Los datos de millones de individuos recolectados, interpretados y segmentados atendiendo a diversos factores, como los psicológicos-personalidad en el escándalo de Cambridge Analytica, han ayudado a sus usuarios a ganar elecciones y referéndums⁴.

Las opiniones de los expertos han perdido gran parte de su prestigio. Desde hace décadas, podemos encontrar que expertos competentes y con autoridad en el tema ofrecen diagnósticos y tratamientos diametralmente opuestos partiendo de los mismos hechos. Ante esta situación, la tentación de encontrar una estimación objetiva, alejada del ruido, aséptica y técnica, basada solo en números, gana adeptos; en la tecnología pura y sin mácula estaría la solución. Así se explica la tajante afirmación («todo lo que no se mide, no me interesa») que abre este artículo y que también podemos relacionar con la facilidad para evaluar los resultados y actuar en consecuencia.

Decisiones racionales y... económicas

Una de las principales aplicaciones de la inteligencia es que permite anticiparse en el tiempo para que el decisor pueda tomar mejores decisiones. En consecuencia, Heuer y Pherson establecen la función «apoyo a la decisión», donde engloban métodos como el análisis DAFO, matriz de decisión, etc.⁵. Así mismo, en algunos estudios sobre análisis de inteligencia, en el ámbito de la «ayuda a la decisión», se incluyen técnicas como teoría

³ PEIRANO, Marta. *El enemigo conoce el sistema*, Debate, 2019.

⁴ Netflix. *El gran hackeo (o Nada es privado)*, documental sobre Cambridge Analytica, 2019.

⁵ HEUER Richards, J. Jr.; PHERSON, Randolph H. *Técnicas analíticas estructuradas para el análisis de inteligencia*, Plaza y Valdés, 2015, pp. 279-301. Versión en español de la 1.ª edición, de 2010, de este libro. Disponible en: <http://www.plazayvaldes.es/libro/tecnicas-analiticas-estructuradas-para-el-analisis-de-inteligencia/1493/>

de juegos o ponderación de criterios (AHP, proceso analítico jerárquico, método Electre, etc.), que «estrictamente» no corresponden a inteligencia; los factores son conocidos y no existe tanta incertidumbre, pero ayudan a tomar decisiones racionales.

Podemos pensar que la principal razón para crear algoritmos es porque cumplen esa función: tomar decisiones objetivas apoyadas en datos, evitando decisiones subjetivas de individuos con criterios distintos. Sin embargo, esto es así solo a medias, ya que con mucha frecuencia el algoritmo se crea porque previamente ha habido una reducción de plantilla, era necesario optimizar los recursos que todavía quedaban y abaratar costes. Esto se conseguía mediante una adecuada planificación y asignación de recursos.

De hecho, su antecedente más cercano lo constituye la «investigación operativa» (también denominada *investigación de operaciones* o *análisis operacional*), una rama de las matemáticas que tuvo gran auge en la Segunda Guerra Mundial, donde los aliados la utilizaron para lograr la máxima eficiencia en el uso de sus recursos. Dado el éxito alcanzado, después de la guerra fue ampliamente utilizada en logística y planificación.

Muchos algoritmos constituyen una primera fase de decisión, una mecánica y relativamente fácil, muchas veces selección de trazo grueso, quedando otra fase a cargo de personas, especialistas en la materia en cuestión, que afinan la solución final⁶. Si los recortes han sido tan importantes que estas dos fases no existen y todo lo hace el algoritmo, el riesgo de error se incrementa.

Por esta misma razón de optimización de recursos, un algoritmo tendrá una versión inicial y actualizaciones continuas. Muchas empresas se juegan su futuro en que todo el proceso vaya bien y puedan ser mejores que su competencia; la retroalimentación y el perfeccionamiento han de ser constantes.

⁶ OLÍAS, Laura. «Cuando el inspector de trabajo es un algoritmo: avances y temores sobre las sanciones laborales automatizadas», *eldiario.es*, 29 de enero de 2021. Disponible en: https://www.eldiario.es/economia/pasa-si-inspector-trabajo-robot-avances-temores-sanciones-automatizadas-aprobado-gobierno_1_7177360.html

Algoritmos, un breve repaso: qué son, qué hacen y cómo lo hacen, cuánto tardan en hacerlo

Para la Real Academia Española, un algoritmo es un «conjunto ordenado y finito de operaciones que permite hallar la solución de un problema». Como en la práctica se realizan informáticamente, se han convertido en sinónimo de software o programa informático. Pero el creador del *like* en Facebook va más allá: los algoritmos complejos pueden ser ya considerados inteligencia artificial⁷, y, desde luego, aprenden automáticamente (*machine learning*). Sin duda son otra categoría y «compiten» con nuestro cerebro de *Homo sapiens* y con nuestras debilidades, tan humanas también. No necesitan superar nuestra inteligencia y nuestras fortalezas, les basta con aprovechar nuestras flaquezas.

Hay algoritmos de muchos tipos. Algunas de sus aplicaciones más usuales son: ordenar o evaluar, haciendo un *ranking* de: universidades; ciudades ecológicas, *friendly* o con más calidad de vida⁸; seleccionar, recomendar o predecir el éxito (noticias, música, series); analizar riesgos (de reincidencia, de impago, de salud, de accidente, etc.); planificar suministros u horarios; optimizar rutas de reparto o patrullaje para prevenir robos y crímenes o reflejar puntos calientes (delitos) en una zona⁹.

También los hay sobre rendimiento deportivo, en béisbol y baloncesto; serían los denominados como PECOTA y CARMELO, respectivamente¹⁰. En ellos se realizan proyecciones para predecir el rendimiento de los jugadores del presente, comparándolos con las estadísticas de otros jugadores «similares» del pasado y, en consecuencia, adelantarse a los rivales y fichar talentos prometedores. Y, cómo no, también en el fútbol¹¹, aunque la fortuna de estos algoritmos es desigual.

⁷ ROSENTEIN, Justin. Entrevista minuto 47 en *El dilema de las redes sociales*, documental disponible en Netflix, 2020.

⁸ PLAZA, Analía. «Qué hay detrás de los rankings de “mejores ciudades para teletrabajar” en los que España sale tan bien parada», *eldiario.es*, 8 de diciembre de 2020. Disponible en: https://www.eldiario.es/economia/hay-detras-rankings-mejores-ciudades-teletrabajar-espana-sale-parada_1_6481578.html

⁹ PREDPOL. *Predictive Policing Company*. Disponible en: <https://www.predpol.com>

¹⁰ TOBOSO, Fernando. *Grado de acierto de estimaciones y pronósticos. criterios de evaluación de la metodología y de la calidad de los análisis*, Instituto Español de Estudios Estratégicos, 19 de julio de 2016. Disponible en: http://www.ieeee.es/Galerias/fichero/docs_marco/2016/DIEEEM12-2016_Metodologia_CalidadAnalisis_FernandoToboso.pdf

¹¹ DRIBLAB. Empresa consultora de Big Data para profesionales del fútbol. Disponible en: <https://www.driblab.com/es/>

Desde hace años son una realidad, siendo su implantación masiva, también en España. Abarcan aspectos insospechados, desde ofrecer resultados diferentes de una misma búsqueda de Google (personalizados a cada usuario en su ordenador, «optimizados» según su historial de búsqueda) hasta determinar el tiempo de espera al teléfono en un servicio de atención al cliente (en función de lo «buen» cliente que seamos) o destapar el fraude fiscal¹².

En palabras de una premiada científica de datos: «Primeramente, se presenta el problema o reto del mundo “real”, se identifica cuáles son las características claves y, finalmente, se transforma en un problema estadístico empleando algoritmos. Se puede obtener la solución en una semana, un mes o un año dependiendo de la complejidad del problema, aunque a veces no existe una solución»¹³.

Si tenemos los datos y sabemos lo que queremos, el software dará con una primera solución en márgenes medidos en horas o días. Requerirá de su comprobación y refinado en nuevas versiones (semanas, meses), en un proceso de perfeccionamiento que nunca se detendrá, pues tendrá una traducción directa en beneficio económico. Una vez recopilados los datos, el acortamiento increíble de los plazos en los que el algoritmo proporciona la solución es otra ventaja que facilita su implantación y difusión, además de su bajo coste y su acierto.

Otro factor que explica su éxito, sobre todo de los algoritmos empleados en redes sociales, es la incorporación de técnicas psicológicas para mantener la atención del usuario e incluso su adicción. Mecanismos teorizados desde 1948 por B. F. Skinner (reclamo, acción, recompensa, refuerzo de intervalo variable) y desde 1998 por B. J. Fogg (motivación, habilidad y señal), que ya mostraron su eficacia en las máquinas tragaperras y sus palancas¹⁴.

En resumen, el algoritmo es una herramienta científica de primer orden, ya que: describe un fenómeno, descubre regularidades y patrones y predice comportamientos futuros.

¹² GONZÁLEZ, Ignacio; MATEOS, Alfonso, «¿Se puede destapar el fraude fiscal con algoritmos?», *Observatorio Social de «La Caixa»*, octubre 2020. Disponible en: <https://observatoriosociallacaixa.org/es/-/se-puede-destapar-el-fraude-fiscal-con-algoritmos>

¹³ CORDERO, Sara. «Fátima El Baghdadi, ganadora de los Global Awards 2020», *El plural*, 1 de enero de 2021. Disponible en: https://www.elplural.com/economia/empresas/fatima-baghdadi-ganadora-globant-awards-2020-no-genio-persona-disciplinada-esfuerza-objetivos_256142102

¹⁴ PEIRANO, *op. cit.*, capítulo 1.

La importancia de los valores sustitutivos en los algoritmos

Puede existir una gran dificultad en captar un concepto cualitativo de la vida «real» (seguridad, calidad, prestigio, excelencia, *friendly*, sano, solvente, etc.) y traducirlo a características que se puedan medir (variables estadísticas). Si se consigue, habremos capturado su esencia y se podrá replicar y reproducir un modelo de éxito, obteniendo también ese éxito buscado.

En muchas ocasiones, la noción abstracta (riesgo de impago de crédito, por ejemplo) sí está suficientemente demostrado que depende de determinadas variables, que, en palabras de la doctora en matemáticas por Harvard Cathy O'Neil¹⁵, serían denominados «valores relevantes» (en el ejemplo anterior: ingresos, importe de la deuda e historial de pago de facturas), pero pueden existir limitaciones legales sobre su uso (consentimiento del individuo, leyes de protección de datos, prohibición de uso en marketing). Tampoco pueden obtenerse datos que impliquen una discriminación por razón de sexo, raza, religión, etc. Así mismo, puede ocurrir que sean fenómenos nuevos o únicos y haya escasa información de la que partir.

En estos casos, los diseñadores del algoritmo buscan otras variables sucedáneas (los valores sustitutivos o *proxies*), aunque no esté suficientemente demostrada la influencia de estas nuevas variables en la definición del concepto cualitativo, pero son datos sobre los que no hay limitaciones sobre su obtención y gestión, el paraíso para las nuevas empresas tecnológicas, que no sufren las regulaciones de otras grandes empresas, como los gigantes de las telecomunicaciones. Así, en Estados Unidos es ilegal utilizar test de inteligencia para contratar a una persona, por lo que son sustituidos por... test de personalidad, aunque estos no se hagan con todas las garantías (*test online*) y no midan lo mismo.

Uno de los muchos ejemplos, y representativo, del uso de un algoritmo con valores sustitutivos en lugar de valores relevantes, que recoge la doctora O'Neil¹⁶, sería el utilizado por compañías de seguros norteamericanas para calcular el precio del seguro del automóvil, donde, en lugar de fijar la tarifa en función del historial del conductor (multas, accidentes, etc.), lo hace en función de sus factores socioeconómicos y patrones de consumo (y los de sus amigos y vecinos, por ser supuestamente similares). Esto tiene

¹⁵ O'NEIL, Cathy. *Armas de destrucción matemática*, Capitán Swing, 2017.

¹⁶ *Ibidem*, p. 206; «The truth about car insurance», *Consumer Reports*, 30 de julio de 2015. Disponible en: [Car Insurance & Auto Insurance Special Report - Consumer Reports](#)

una traducción en el precio del seguro, que perjudica a muchos conductores, aunque no sean los peores al volante.

Igualmente, para «el Gran Hermano», el antaño importante «¿qué hacemos?» ha sido sustituido por el mucho más asequible *proxy* «¿dónde estamos?», a disposición de las compañías telefónicas y Google (y, a través de él, de otras terceras partes). De igual forma, el contenido de nuestros correos electrónicos ha sido reemplazado por sus metadatos.

Como minusvaloramos nuestra privacidad y la importancia de nuestros propios datos, los cedemos despreocupados o resignados para conseguir el acceso o el uso de programas atractivos. El resultado es que datos no faltan, apareciendo en los sitios y para los propósitos más inusitados. Pueden seleccionarse decenas, cientos o miles de un mismo individuo u objeto, asumiéndose que esa gran cantidad permite una mejor representación de la calidad buscada; se aproximarán y, con las nuevas versiones del algoritmo, se perfeccionarán.

Los valores sustitutivos ofrecen una aproximación a la abstracción que definen, por lo que se busca, además de su fácil observación, su similitud. Lógicamente, cuanto más correlación exista entre ellos y el concepto cualitativo, mejor lo encarnará. En ocasiones puede no existir una correlación evidente, y solo el tiempo determinará si estamos ante valores relevantes o sustitutivos. Sin embargo, la mayor parte de las veces no existe esa duda y se opta conscientemente por unos u otros en función de su disponibilidad o los efectos que se persigan. De esta forma, se elegirán los valores sustitutivos que más favorezcan el fin deseado.

Estas aproximaciones constituyen una referencia inicial que ahorra errores de bulto; ofrecen un punto de partida. Su acierto en la sustitución será dispar y, aparte de consideraciones sobre la finalidad perseguida, distinguirá a un buen analista, ya que puede no ser obvia la relación entre esos valores y la noción que debe representar.

La gestión de todos estos datos sería sencilla debida a los avances informáticos, impensables hace unas décadas. El *big data* lleva ya años entre nosotros y se concreta en estos algoritmos con los que se codifica el pasado y se ofrece una previsible visión del futuro.

Por último, este proceso se valida. Se confirma que hemos capturado ese concepto por la respuesta mediática, o social; el éxito académico, económico, deportivo, o en el ámbito en el que se inscriba ese concepto.

Si realizásemos una comparación entre los valores relevantes, los valores sustitutivos y su empleo en algoritmos, su resumen sería el siguiente:

	Valores relevantes	Valores sustitutivos
Correlación con concepto a estudiar	Fuerte y demostrada.	«Suficiente».
Finalidad	Proporcionar solución/ones al problema planteado.	Proporcionar solución/ones al problema planteado. Frecuentemente, favorecer una opción determinada.
Cantidad de valores	Limitada, en torno a la decena.	Muy numerosa, pueden ser cientos o miles.
Acceso y gestión de los valores	Regulado y protegido.	Sin regular, vendidos o cedidos interminablemente.
Extracción temporal	Datos del pasado.	Datos del pasado.
Transparencia	Se suelen difundir los tenidos en cuenta.	No se difunden por completo.

Tabla 1. Comparación entre valores relevantes, sustitutivos y su empleo en algoritmos. Fuente.

Elaboración propia.

Aunque los valores sustitutivos no reflejen la realidad, sí que nos deslumbran con su brillante envoltorio tecnológico.

Estos *proxies* utilizados en procesos cuantitativos serían asimilables a las analogías en procesos cualitativos. Los historiadores, por ejemplo, recurren frecuentemente a las analogías confrontando situaciones del presente con situaciones del pasado, comparando y contextualizando sus semejanzas y diferencias, para facilitar la explicación del hecho actual y su posible evolución.

Hay, por tanto, una búsqueda de valores conocidos o situaciones del pasado que nos puedan ser útiles en el presente para el futuro próximo.

De la cualidad abstracta a las variables cuantitativas

Tomando como base el magnífico libro *Armas de destrucción matemática*, de la doctora O'Neil¹⁷, gran especialista en la materia, se desarrollará como ejemplo un algoritmo de los más antiguos y representativos, el que establece el *ranking* de las mejores universidades norteamericanas. Se supone que evaluará la noción «calidad y prestigio universitario», siendo el modelo deseado, a imitar (con los requisitos a cumplir) el de las universidades de más renombre, como Yale, Harvard, Stanford o el MIT, que ya cuentan con gran prestigio internacional, han dado con «la fórmula del éxito» que ahora se pretende captar y replicar.

Esta tarea se la planteó un magacín en horas bajas (*US. News & World Report*), que deseaba aumentar su audiencia, evaluando nada menos que a 1800 facultades y universidades. En 1988, su listado vio la luz. Un 75 % estaba elaborado con un algoritmo y el resto en base a opiniones de empleados universitarios. Los resultados parecían sensatos y no levantaron polémica, pero en años sucesivos la lista se convirtió en referencia sacrosanta en todo el país, produciendo un bucle que se realimentaba, premiaba a los bien clasificados y penalizaba a los demás. Nadie podía imaginar la acogida y éxito que obtuvieron, de tal forma que la revista desapareció en 2010, pero los *rankings* de todo tipo se extendieron, debido a la gran aceptación y credibilidad con la que fueron recibidos.

La manera usual de proceder es descomponer ese complejo tema en muchos temas (o preguntas/variables) más sencillos que puedan ser observados (y traducidos a cantidades objetivas). Esa cantidad no sería problemática de gestionar y, teóricamente, nos aproximaría más a entender la esencia del concepto estudiado.

Así, de las decenas o cientos de variables posibles se mencionarán algunas como ejemplo:

Variable 1: Precio de la matrícula

Variable 2: Precio de los créditos

Variable 3: Ubicación en zona rural o urbana (ciudad pequeña, mediana, grande, previamente definidas)

¹⁷ O'NEIL, *op. cit.*, pp. 65-86.

Variable 4: Número de solicitudes de ingreso recibidas por cada plaza disponible

Variable 5: Porcentaje de alumnos que aprueban primer curso

Variable 6: Porcentaje de alumnos que se gradúan

Variable 7: Número de alumnos por clase en asignaturas de ciencias

Variable 8: Número de alumnos por clase en asignaturas de letras

Variable 9: Número de horas de prácticas por asignatura

Variable 10: Número de horas dedicadas a conferencias o seminarios por asignatura

Variable 11: Número de conferencias impartidas por premios nobel

Variable 12: Número de profesores de la universidad que han obtenido un Premio Nobel

Variable 13: Número de artículos de profesores en revistas de prestigio (previamente identificadas)

Variable 14: Número de citas de esos artículos

Variable 15: Facilidad de acceso para personas con movilidad reducida

Variable 16: Número de habitaciones y alojamientos en el propio campus

Variable 17: Número de habitaciones y alojamientos a menos de 10 km del campus

Variable 18: Número de habitaciones y alojamientos entre 10 y 30 km del campus

Variable 19: Cantidades donadas por exalumnos a la universidad

Variable 20: Cantidades donadas por empresas y fundaciones privadas a la universidad

Variable 21: Porcentaje de exalumnos que han encontrado trabajo nueve meses después de graduarse

Atendiendo a las distintas facetas, relativas a profesorado, alumnado, instalaciones (deportivas, clases, laboratorios, bibliotecas y salas de estudio, alojamiento, comedores, cafeterías, etc.), podríamos llegar sin mucho esfuerzo al centenar de variables para captar la excelencia de esas universidades de la «primera liga» norteamericana (hay otros cientos de ellas que ni entrarían en el *ranking*).

Un aspecto importante sería cómo se obtienen las respuestas a esas variables, teniendo en cuenta el gran volumen de universidades valoradas, trabajo ingente para los redactores de una revista en apuros. Algunas de ellas son de acceso público y podrían ser obtenidas por dichos redactores (como el precio de la matrícula, pero en 1988 no a través de Internet), otras solo pueden ser contestadas por la propia universidad, como el

«porcentaje de exalumnos que han encontrado trabajo nueve meses después de graduarse». A este respecto, ¿cómo lo investigaba la universidad? En algunos casos escribía una carta al exalumno y, si no recibía respuesta, interpretaba que estaba muy ocupado porque ya tenía trabajo. En la mayoría de los casos no sabemos cómo lo averiguaba.

Como este *ranking* lleva haciéndose más de treinta años, podemos ser benévolos y pensar que algunas variables y algunas universidades han recibido una auditoría, realizándose una investigación más exigente a sus respuestas, pero cuando la única opción posible es obtener ciertos datos para la evaluación a través del propio evaluado, se corre un riesgo sobre la objetividad de dichos datos.

Sean los redactores los que consiguen las respuestas a las variables o sean las universidades las que contestan un cuestionario, son los propios periodistas, un grupo de expertos o los programadores del algoritmo los que deciden qué variables incluir y cuáles no, y el peso que otorgan a cada una de ellas. En este caso, la decisión correspondió a los periodistas, que no incluyeron variables como el coste de la matrícula y las tasas universitarias, ya que, de incluirlas, las universidades que teóricamente servían de modelo (Yale, etc.) no saldrían en los primeros lugares. En otras palabras, el diseño del algoritmo, al elegir y priorizar unos datos y desechar otros, permite orientar sus resultados en una dirección determinada.

El algoritmo es un paquete cerrado, opaco e inapelable. No admite explicaciones ni reclamaciones, «son datos», ellos y todo lo que lo rodea, sería perfecto por definición, ajenos a opiniones subjetivas y revestidos del aura de autoridad de la «ciencia».

Con este *ranking*, los alumnos y sus familias, aparentemente, sabrían a qué atenerse, decidir racionalmente y escoger la mejor universidad.

La universidad pobremente evaluada, sin embargo, solo podría imaginar un porcentaje de los factores evaluados, no el resto, por lo que, para mejorar su puesto en el *ranking*, a lo sumo intentaría mejorar ciertos aspectos, dando palos de ciego, para ver su influencia en el listado del año siguiente. De esta forma, su objetivo, más que la mejora académica, sería contentar al algoritmo, ya que de su puesto en la lista puede depender su posibilidad de captar alumnos y donaciones, y, en definitiva, su supervivencia.

Ante los inconvenientes de este *ranking* u otros similares, ¿cuál sería una posible solución? Nos la comenta la doctora O'Neil: «En lugar de crear un nuevo *ranking*, el

Departamento de Educación estadounidense publicó montones de datos en un sitio web. Y ahora los estudiantes pueden hacer sus propias preguntas sobre lo que les interesa a ellos, como la ratio de alumnos, la tasa de licenciados y la deuda media de los estudiantes al terminar sus estudios. No necesitan tener conocimientos de estadística ni saber la ponderación de las distintas variables. El propio software, al igual que hacen los sitios web de las agencias de viajes en línea, crea modelos individuales para cada persona»¹⁸.

Posiblemente el estudiante filtre la búsqueda atendiendo a una decena de criterios, los más importantes para él, puesto que sabe en qué valores relevantes reside la excelencia que busca en la universidad; cambiará gustoso decenas de criterios por un número muy limitado de ellos. Afortunadamente, cada vez más webs ofrecen esta alternativa adaptada al usuario, que le permiten ser el que elija y decida.

Valores sustitutivos y superpronosticadores

El psicólogo norteamericano Philip Tetlock, en su libro *Superpronosticadores*¹⁹, también explica y destaca la importancia de ciertos valores, asimilables a los valores sustitutivos vistos anteriormente, que ayudan a ofrecer respuestas concretas.

En él continúa su trabajo de investigación iniciado en 1984 sobre el juicio político de los expertos. En este Good Judgment Project sigue planteando preguntas sobre acontecimientos futuros (en márgenes medidos en meses), que solo permiten respuestas cuantitativas, cifras o porcentajes que facilitan su evaluación y descubren expertos en realizar pronósticos probabilísticos: los superpronosticadores.

Así, a principios de 2015 plantea una de sus características preguntas: «¿qué probabilidad de atentados yihadistas hay en ocho países europeos (que se detallan), del 21 de enero al 31 marzo de 2015?»²⁰. Con posterioridad analiza la respuesta del superpronosticador que más se acercó a la realidad. Este experto, primeramente, buscó una referencia inicial en sucesos similares de la que partir para evitar errores de bulto; después la modificó atendiendo a las circunstancias particulares y concretas de esos

¹⁸ O'NEIL, *op. cit.*, p. 86.

¹⁹ TETLOCK, Philip. *Superforecasting*, Crown Publishing Group, 2015. [Edición en español: *Superpronosticadores*, Katz Editores, 2017]

²⁰ *Ibidem*, p. 126.

primeros meses de 2015. Esa primera referencia la buscó en el pasado, en un dato sólido, objetivo y conocido: el número de atentados yihadistas en esos ocho países en los últimos cinco años, lo que constituye el promedio de atentados por año; en estadística sería la tasa base. Este acierto o error inicial en encontrar un valor «similar» del pasado es clave y distingue al superpronosticador, pues se irá arrastrando y condicionará notablemente el acierto final.

Antes de hacer definitiva su estimación, el superpronosticador la difundió en un foro interno del proyecto, para someterla a la revisión por sus pares de pronosticadores, pudiendo corregir su estimación hasta la fecha inmediata a la pregunta, aunque todo quedase registrado.

Este procedimiento, y en este orden general-particular (valor de referencia extraído del pasado, corrección con circunstancias específicas del presente), se repite en otros ejemplos en el libro. Tetlock lo define de acuerdo con Daniel Kahneman (psicólogo y premio nobel de Economía por sus aportaciones respecto al juicio humano y la toma de decisiones bajo incertidumbre), con su «visión desde afuera» y posterior «visión desde adentro»²¹, y también en consonancia con el «pensamiento de fuera-adentro» de Heuer y Pherson²².

Tetlock también llama la atención sobre nuestra tendencia natural a reemplazar, incluso inadvertidamente, una pregunta complicada por otra mucho más sencilla, denominándolo *sustitución de atributo*²³.

Todo lo anterior enlaza con lo visto anteriormente en la búsqueda de valores sustitutivos y también con la forma de pensar de Enrico Fermi, que Tetlock y su equipo admiran, detallando cómo Fermi estimaba el número de afinadores de piano en Chicago²⁴. Este gran físico resolvía una cuestión muy complicada de la que apenas tenía datos mediante cuestiones más sencillas que le eran más fácil de valorar, en una secuencia lógica que disminuía el margen de error. Seguramente no acertaba con la cifra exacta, pero sí con su magnitud.

²¹ KAHNEMAN, Daniel. *Thinking, Fast and Slow*, Nueva York: Farrar, Straus and Giroux 2011. [Edición en español: *Pensar rápido, pensar despacio*, Editorial Debate, 2012]

²² HEUER, *op. cit.*, p. 213.

²³ TETLOCK, *op. cit.*, p. 47.

²⁴ MONTEJANO, Javier. «Preguntar con éxito: el problema de Fermi», *Club-mba*, 22 de mayo de 2015. Disponible en: <https://www.club-mba.com/2015/05/22/preguntar-con-exito-el-problema-de-fermi/>

Conclusiones

No se intenta reconvertir a nadie en científico de datos, tarea imposible para quien no tenga una formación adecuada, sino:

- Comprender el valor transcendental de los datos: su proceso de obtención y selección condiciona y sesga las soluciones finales que ofrecen los algoritmos, mecanismos omnipresentes y muy influyentes en la vida cotidiana.
- Resaltar la importancia de los valores relevantes, los valores sustitutivos y otras técnicas similares ampliamente utilizadas, muchas veces la única alternativa asequible a pesar de sus inconvenientes, pero toda una forma de pensar para resolver cuestiones complejas y abstractas mediante procesos más simples y concretos.
- Cabe recordar que los resultados o las estimaciones que ofrecen tanto los algoritmos como los superpronosticadores no son indiscutibles e inamovibles. Para ser exitosas deben ser continuamente revisadas y perfeccionadas.

*Fernando Toboso Marqués**

Teniente coronel (R) del Ejército de Tierra
Profesor de Metodología de Análisis en CISDE