ECONOMÍA DEL DATO: LUCES Y SOMBRAS

DAVID RÍOS INSUA

ICMAT-CSIC

Un rasgo global de esta última década está siendo el rápido crecimiento en las capacidades de muchas organizaciones para explotar los avances en tecnologías de la información, la modelización estadística y la investigación operativa, para capturar y procesar datos sociales, de mercado y de operaciones y apoyar su toma de decisiones. Cada vez más, la información obtenida de datos constituya un ingrediente esencial que posibilita procesos

más automatizados y servicios y productos más personalizados, dando lugar a la *Economía del Dato* (Brynjolfsson y Kahin, 2002).

Como resultado, la analítica de negocios se ha convertido en un campo floreciente para la consultoría empresarial. Sin embargo, aunque muchas decisiones de algunos gobiernos a menudo vienen apoyadas con métodos tradicionales del análisis de políticas públicas, como el análisis de coste-beneficio, pocos departamentos y agencias gubernamentales están aprovechando de forma sistemática las grandes masas de datos disponibles y los métodos avanzados de la estadística y del aprendizaje automático para obtener evidencias que informen sus decisiones. En la industria, el énfasis se ha puesto en resolver problemas relacionados con el análisis de datos masivos y complejos, en entornos que suelen denominarse de Big Data (Rios Insua y Gómez-Ullate, 2019). Nos enfrentamos pues a una nueva era en la que hay una enorme cantidad de datos digitalizados sobre numerosos temas de interés potencial para una empresa o un gobierno. Sin embargo, con bastante frecuencia, dicha información es altamente desestructurada y difícil de gestionar y, no inusualmente, poco relevante, por aportar poco valor.

El análisis de estos datos no estructurados, o no muestreados, es un reto mayor que ha dado lugar a nuevos paradigmas como la Ciencia y la Ingeniería de Datos. Los datos no estructurados se caracterizan por que su formato es muy variable y no pueden almacenarse en bases de datos relacionales tradicionales sin un esfuerzo significativo que conlleve transformaciones complejas. Se emplean así bases de datos NoSQL, más escalables. Su gestión requiere marcos computacionales que permitan realizar cálculos basados en el procesamiento distribuido sobre conjuntos más pequeños. Finalmente, se necesitan también infraestructuras de almacenamiento que posibiliten el resumen y análisis de los datos. Además de los avances tecnológicos, también hay nuevas clases de métodos de análisis que permiten la extracción de información: aparte de técnicas

tradicionales como modelos de regresión o clasificadores de *k*-vecinos más cercanos, se dispone de métodos más recientes como los árboles de clasificación y de regresión, las máquinas de soporte vectorial o, muy especialmente, las redes neuronales profundas en sus distintas versiones.

En cualquier caso, los datos parecen ahora más accesibles a los gestores, que tienen una gran oportunidad para tomar mejor sus decisiones para aumentar sus ingresos, reducir sus costes, mejorar el diseño de sus productos, detectar y prevenir el fraude, o la mejora de la conversión de clientes. Esto ha conducido a un nuevo concepto de organización que toma decisiones basadas en la evidencia, con ejemplos claros como Alphabet, Meta, Walmart, Alibaba o algunas de las líneas aéreas más avanzadas.

Nuestro objetivo aquí es ilustrar a través de ejemplos el enorme potencial (luces) de estas metodologías y tecnologías para resolver problemas de carácter social, además de mostrar algunas cuestiones preocupantes (sombras) que deben afrontarse para su adecuada utilización, concluyendo con algunas recomendaciones de futuro relevantes en la Economía del Dato.

ALGUNAS LUCES EN LA ECONOMÍA DEL DATO

Comenzamos presentando algunos ejemplos que muestran el enorme potencial de las tecnologías y metodologías relacionadas con el Big Data y la Inteligencia Artificial a la hora de resolver problemas de tipo social. Todos ellos se refieren a proyectos recientes en los que hemos estado involucrados con un impacto positivo en la sociedad.

Colaboraciones B2G en datos para el beneficio social

Esta primera luz de la Economía del Dato, muestra cómo los tenedores privados de datos pueden contribuir al mejor desarrollo de la sociedad, facilitando el intercambio de datos B2G. (1) En concreto, se describe un proyecto que aprovecha métodos de la Inteligencia Artificial para construir a partir de una gran base de datos un sistema de que facilita la toma de decisiones en políticas de salud pública.

Las enfermedades cardiovasculares (ECV) son la causa principal de mortalidad en Europa (45% de todas las muertes), con costes anuales asociados estimados en 210 billones de euros (Wilkins *et al.*, 2017). Describimos aquí brevemente cómo el tratamiento de grandes bases de datos con métodos de la IA permite realizar contribuciones relevantes que mejoran el entendimiento y, muy especialmente, el tratamiento de la enfermedad cardiovascular. El modelo descrito y su implementación está disponible en Rios Insua *et al*. (2021b).

Los datos provienen de los reconocimientos médicos anuales de trabajadores afiliados con una compañía privada de seguros, convenientemente anonimizados y securizados. Se complementan con información del censo, basada en el código postal de vivienda de los individuos, a partir del cual obtenemos de bases de datos públicas, el status socio-económico y el nivel educativo medio correspondiente. Finalmente, un intensísimo trabajo de depuración de outliers, duplicados, datos mal registrados, valores faltantes y transformación de variables, proporcionó un conjunto de datos estructurado y completo sobre el que emplear modelos estadísticos y de aprendizaje automático. Creamos así, a partir de un conjunto crudo de datos (con casi cinco millones de casos y 40 variables) mal formado e incompleto, una base bien estructurada y limpia. De hecho, se convierte en una base de datos probabilística en forma de red bayesiana (Castillo *et al.*, 2012), cuya estructura y tablas de probabilidad en nodos se aprenden a partir de los datos. Esta aproximación posibilita predicciones para cada grupo de factores a partir de las capacidades de la red para realizar inferencia probabilística.

Este proyecto, en particular, pone el énfasis en la evaluación predictiva de los distintos factores de riesgo CV, y en especial en la actividad física, integrando factores como *depresión*, *duración del sueño* y *status socioeconómico*. Específicamente, las variables que finalmente se incluyeron en el modelo fueron: *Factores de riesgo cardiovascular (FRCV) no modificables* (Sexo, Edad, Nivel educativo, Nivel socioeconómico); *FRCVs modificables* (Indice de masa corporal, Actividad física, Duración del sueño, Historial fumador, Ansiedad, Depresión); y *Condiciones médicas* (Hipertensión, Hipercolesterolemia, Diabetes). Para construir la red, empleamos un proceso en dos pasos. En el primero, se usó un algoritmo de búsqueda bayesiana para aprender una estructura inicial a partir de los datos. En el segundo, se implementó un proceso iterativo en el que se mostraba a expertos diversas estructuras pidiéndoseles eliminar o añadir arcos relevantes razonando a partir de un mecanismo de inferencia referido a nodos progenitores y nodos hijos. Una vez con la estructura se asignaron las tablas de probabilidad de los nodos con modelos multinomial-Dirichlet, partiendo de distribuciones a priori uniformes (French y Rios Insua, 2000).

Basados en esta red, se pueden implementar numerosos usos en terapia y políticas de salud pública, aprovechando su capacidad para incorporar información, mediciones y observaciones y propagarlas a través de la red y modificar las distribuciones en los otros nodos empleando, tal vez varias veces, la fórmula de Bayes (Nielsen y Jensen, 2008). Así, se pueden determinar las probabilidades de las enfermedades dadas ciertas condiciones; se pueden también hacer hipótesis sobre evidencias relacionadas con varias cuestiones referidas a salud. Además, una vez determinado un caso de interés (grupo o individuo) y evaluada la probabilidad correspondiente, podemos encontrar los hechos influyentes sobre tal afirmación. Esto es especialmente relevante para los FRCV modificables para los que podemos explo-

rar la mejor modificación y sugerirla al grupo o individuo de interés.

Recuérdese, sin embargo, que para decidir la mejor recomendación necesitaríamos tener en cuenta los posibles impactos de las condiciones médicas y los tratamientos a través de funciones de utilidad y utilidades esperadas, como ilustraremos en los restantes ejemplos.

Apoyo a la toma de decisiones estratégicas públicas

La luz aquí presentada se refiere a aprovechar la evidencia aportada por datos e incorporarla a un sistema de ayuda la decisión que facilita la asignación de recursos estratégicos en una agencia estatal, con enormes ahorros de costes operativos. El modelo descrito y su implementación se describe en detalle en Elvira *et al*. (2020).

La Organización de Aviación Civil Internacional (OACI) persigue que la aviación sea el modo de transporte más seguro por ser un factor clave en el desarrollo sostenible de las naciones. Globalmente, la aviación aérea opera a un nivel de seguridad muy alto. Por ejemplo, en Europa, la tasa media de accidentes fatales fue de 1.3 por cada 10 millones de vuelos. Sin embargo, las agencias estatales tratan de mejorar permanentemente este complejo sistema. Para ello, los países deben desarrollar un Plan Estatal de Seguridad Aérea (SSP) para la aviación civil, que debe incluir los objetivos nacionales de seguridad y afecta a todas las partes interesadas (autoridades, proveedores de servicio, aeropuertos,...). Los planes deben identificar las fuentes principales de inseguridad y el conjunto de acciones para mitigar y controlar los riesgos asociados a las mismas.

Hasta hace relativamente poco tiempo, la gestión de riesgos en seguridad aérea se ha basado en el uso matrices de riesgos, a pesar de tener varios fallos bien conocidos Cox (2008). Para superarlos, la Agencia Estatal de Seguridad Aérea (AESA) desarrolló una metodología más rigurosa que facilita la asignación óptima de recursos, haciendo que los sucesos que comprometen la seguridad aérea sean menos frecuentes y, en caso de que se materialicen, sean menos dañinas. En su contexto, AESA gestiona 86 tipos de sucesos de seguridad (desde salidas de pista hasta fallos de motor, pasando por colisiones en tierra) con cinco niveles de severidad. Un análisis previo, que incluyó una revisión de la literatura y una tormenta de ideas con altos ejecutivos de AESA, condujo a la identificación de ocho consecuencias relevantes para la seguridad operacional de la aviación en nuestro país: (1) muertes asociadas al funcionamiento del sistema de aviación; (2) lesiones menores y (3) lesiones graves; (4) retrasos y (5) cancelaciones asociadas a los incidentes; (6) operaciones de mantenimiento y (7) reparaciones; y, finalmente, (8) pérdida de imagen-país. De nuevo, además de las excelentes bases de datos disponibles en la organización, se requirieron cruces con bases de datos externas a través de técnicas de web scraping, así como un intenso trabajo de armonización de datos.

El objetivo final de la metodología era encontrar la asignación de recursos que optimizase la seguridad operacional de la aviación nacional, reduciendo así, en la medida de lo posible, las ocurrencias de los distintos tipos de severidad, su gravedad y las consecuencias resultantes. Esto conlleva predecir el impacto de la cartera de recursos de seguridad (principalmente, tiempo de inspección) a implementar sobre la tasa y la gravedad de los distintos tipos de ocurrencia, así como de sus impactos y, después, encontrar la cartera óptima de seguridad, aquella que maximiza la utilidad esperada. Tales actividades conllevaron la construcción de casi dos mil modelos de predicción (no triviales) por lo cual fue necesaria desarrollar un proceso de automatización de determinación y ajuste de modelos, así como desarrollar modelos novedosos de predicción.

Para facilitar la implementación de la metodología introducida, se diseñó el sistema RIMAS. Su valor potencial se verificó comparando los resultados de seguridad reales basados en sus recomendaciones con los que habrían resultado del mantenimiento de las políticas tradicionales de seguridad. El uso de RIMAS proporcionó un rendimiento significativamente mejor en términos de los principales objetivos de gestión, conduciendo a mejoras importantes en la seguridad de la aviación y menores costes de reparaciones, mantenimiento, retrasos y gastos reducidos de aeronaves, estimados en un ahorro anual de unos 800 millones de euros en costes de seguridad equivalentes.

Etica regulatoria

La luz que aquí reflejamos se refiere a cómo incorporar aspectos éticos a las regulaciones introducidas en un sector innovador del máximo impacto social, como es el de los vehículos autónomos (ADS).

Estos van a revolucionar el transporte por carretera (Burns y Shulgan, 2019): facilitado por los avances recientes en IA y en hardware, el transporte masivo mediante vehículos autónomos ha dejado de ser una realidad distante en el tiempo. Sin embargo, la transición a un sistema de circulación totalmente automatizado será un proceso incremental, desde los vehículos conducidos por personas (MVs) a los ADS, como se refleja en la taxonomía de seis niveles de la Society of Automobile Engineers (2018). Numerosas limitaciones relacionadas con su seguridad y robustez operativa probablemente restringirán en la próxima década los ADS a los niveles 3 y 4, que requieren la intervención humana cuando operan fuera de su dominio operativo (ODD) a través de una operación denominada *petición de intervención* (RtI).

Los ADS modernos se basan en una serie de componentes que emplean hardware potente y una variedad de sensores para generar salidas de dirección

y aceleración. La información del entorno se recopila, entre otros, a través de cámaras de luz visible, sensores de detección y rango de luz (LiDAR), o sensores propioceptivos. Una vez recogidos estos datos se utilizan en sistemas que perciben tanto el entorno externo como el interno del ADS. Las entradas de los sensores se utilizan en una secuencia de algoritmos de aprendizaje profundo procesados en plataformas informáticas potentes y compactas diseñadas específicamente para tareas de aprendizaje automático. Los algoritmos se agrupan en tres capas (de percepción, de predicción y de decisión) frecuentemente integradas en una arquitectura de extremo a extremo. Los de la primera capa reciben entradas de los sensores sin procesar; estiman la posición del ADS, así como determinan la geometría y la semántica del vehículo. Para ello, el ADS emplea algoritmos como redes neuronales convolutivas para clasificación (Wu *et al.*, 2017). Sus salidas se emplean como entradas a la capa de predicción que predice cambios en el entorno percibido; empleándose esquemas de modelización que incorporan información incompleta, como procesos de decisión de Markov parcialmente observables (McAllister *et al*., 2017). Los resultados de la capa de predicción entran en la de decisión que se encarga de la planificación de rutas y movimientos, tanto de la ruta macroscópica en la red de carreteras, como de su movimiento granular en el flujo de tráfico, p.ej. véase Claussmann *et al*. (2019). Este esquema general de funcionamiento ha tenido cierto éxito, pero quedan importantes desafíos científicos y tecnológicos por resolver antes de que pueda acaecer la adopción masiva de los ADS en las carreteras.

Para facilitar su integración social, se están implantando diversas iniciativas para aumentar nuestra confianza en los ADSs y facilitar su adopción. El esquema que implementamos en este problema (Rios Insua *et al*., 2021a) hace predicciones del entorno y del estado del conductor y calcula la trayectoria en los siguientes instantes; determina entonces si se alcanzaría mayor utilidad esperada con el conductor o en el modo autónomo y, en función de ello, toma las decisiones correspondientes. Si se invoca una RtI, se acompaña de una evaluación del comportamiento del conductor (DIPA) lo que ayuda a gestionar futuras RtIs (o hacer paradas de emergencia). Este proceso entraña grandes desafíos, algunos de ellos de carácter ético. Así, los modelos y simulaciones ya desarrolladas en el marco de la gestión de RtIs (Rios Insua *et al*., 2021a) han permitido identificar lo que denominamos *dilema fundamental de los ADS de nivel 3 y 4*. Este es un ejemplo paradigmático del tipo de problemas éticos en ADS, de los que son especialmente conocidos los denominados *problemas de tranvía* (Jarvis Thomson, 1985). De forma interesante, el análisis de riesgos permite arrojar luz sobre cómo un ADS puede actuar ante un conflicto ético (Caballero *et al*., 2022), donde se desarrolla un marco basado en la teoría de la decisión, que puede emplearse para gestionar y evaluar la toma de decisiones en ADS. Tiene en cuenta objetivos múltiples: rendimiento del vehículo, duración del viaje, seguridad (la de los pasajeros, las personas en la escena de conducción, el propio vehículo y la infraestructura),... Una vez definidos los objetivos, un productor podría decidir ponderarlos de distinta manera, dando más importancia a aquellos que sean de su interés. Estos pesos se utilizarían para combinar los distintos objetivos en una única función de utilidad que regiría las operaciones del ADS. El interés de este modelo reside en el hecho de que, en caso de accidentes, es posible simular escenarios de conducción múltiples empleando la función de utilidad elegida por el fabricante para guiar las decisiones del ADS. Estas simulaciones permitirían evaluar si el vehículo satisface la regulación vigente, y en caso de no hacerlo, determinar responsabilidades. En definitiva, este modelo conforma un marco que permite evaluar de forma objetiva si las elecciones éticas por parte de fabricantes o usuarios satisfacen las líneas establecidas en una regulación.

Aspectos afectivos en la toma de decisiones

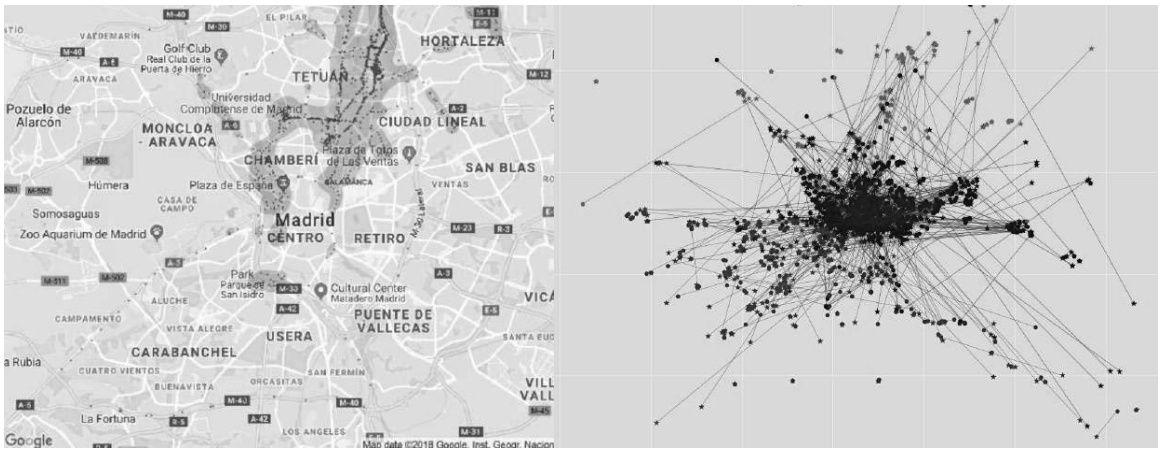
La última luz que esbozamos se refiere a incorporar elementos afectivos en la toma de decisiones para favorecer la conexión con los usuarios. En este caso, el dominio de aplicación sería el de mejora y modernización de la educación.

Continuando con el diseño de agentes que tomen decisiones de forma autónoma, además, ahora deben reflejar o simular emociones, de manera que éstas tengan cierta influencia en su toma de decisiones. Nuestro objetivo funcional último no es meramente descriptivo, sino mejorar la interacción del agente con los usuarios que puedan aparecer en su escena. Nuestro objetivo aplicado es diseñar robots sociales emocionales que puedan adoptar el rol de mentor para un niño o grupo de niños o puedan acompañar, por ejemplo, a personas mayores o pacientes en un hospital.

A los modelos antes descritos, debemos acoplarles conceptos en relación con emociones dentro de lo que se ha denominado *toma de decisiones afectivas* (Loewenstein y Lerner, 2003). Tradicionalmente, las emociones se han considerado como alejadas de la racionalidad. Sin embargo, se han venido produciendo descubrimientos e innovaciones que están modificando la anterior visión tradicional. Por ejemplo, dentro de las neurociencias, se promueve el concepto de inteligencias múltiples (Gardner, 2011) que incluye la inteligencia emocional. Las ideas anteriores se expresan en sistemas computacionales que dan lugar al campo de la *computación afectiva*, véase, p.ej., el trabajo pionero de Picard (1997).

Podemos considerar que, en definitiva, el objetivo final de esta búsqueda es diseñar un agente que sea capaz, dentro de un entorno en el que puede haber otros agentes, de percibir tal entorno y las acciones de tales agentes; en función de estas

FIGURA 1
IZDA: TRAZA DE UN USUARIO DE UNA APP DE PAGO. DCHA: AGRUPACIÓN DE USUARIOS SEGÚN DOMICILIO Y LUGAR DE TRABAJO



Fuente: Elaboración propia

percepciones, mostrar algún tipo de emociones y, finalmente, que éstas se reflejen en la toma de decisiones del agente. Resolver esta tarea requiere que seamos capaces de resolver científica y tecnológicamente una serie de actividades básicas (percepción, inferencia, predicción, afecto, decisión).

La parte más novedosa sería el modelo de preferencias con emociones. Comenzamos identificando los objetivos vitales de nuestro agente. Para ello, nos inspiramos en la pirámide motivacional de Maslow (1943) y construimos objetivos relacionados. Sin pérdida de generalidad, asumimos una función de utilidad aditiva. Los pesos pueden estar ordenados de forma creciente en función de la posición en la jerarquía, para promover que se dediquen más recursos computacionales a los objetivos más básicos. Además, la forma de las funciones componentes de utilidad permite modelizar que, una vez suficientemente satisfechos los niveles en los objetivos más básicos, se pase a perseguir objetivos de nivel superior.

Nos falta enfrentarnos a un concepto elusivo, el de emoción, sobre el que se han dado numerosas definiciones e interpretaciones, véase Russell y Barrett (1999). Nosotros adoptamos una aproximación pragmática, desde el punto de vista computacional, basada en la presencia de emociones básicas, cuya composición da lugar a emociones más complejas. Quedaría entonces definir las emociones a adoptar, cómo se componen y cómo afectan a la toma de decisiones, que pueden verse en detalle en Liu y Rios Insua (2020). Incluyen elementos de personalidad según el modelo HEXACO (Ashton *et al*., 2014); emociones en cuatro grupos (esperadas, inmediatas, referenciales y complejas) y humor y modelos para su definición y actualización.

El esquema anterior se ha implementado en Aiko, una plataforma robótica flexible de Aisoy (https://aisoy.com/), basada en un procesador Raspberry Pi 4, y se ha aplicado con éxito en educación, educación con necesidades especiales, acompañamiento de personas mayores y acompañamiento de personas enfermas.

ALGUNAS SOMBRAS EN LA ECONOMÍA DEL DATO

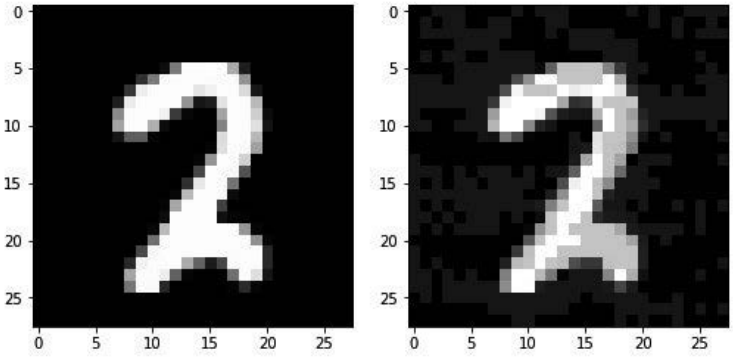
En la sección anterior hemos mostrado, a través de proyectos reales, usos típicos de la Economía del Dato que sugieren su enorme potencial para resolver algunos problemas globales del máximo interés social. Pero, además de estos éxitos, debemos ser también conscientes de los riesgos asociados a tales tecnologías y metodologías, algunos de los cuales describimos a continuación. Proceden también de proyectos recientes. Para justificarlos podríamos apelar a la Declaración Universal de los Derechos Humanos (2), legislación reciente como el Reglamento General de Protección de Datos (RGPD) (3) o el reciente marco ético de la UNESCO sobre valores y principios comunes sobre inteligencia artificial.

Perfilado de individuos

Algunas actividades típicas de la Economía del Dato se refieren al cruce de bases de datos privadas y públicas para crear valor para la sociedad, como ya hemos visto. Sin embargo también conlleva riesgos como es la posibilidad de obtener perfiles demasiado detallados de individuos como mostramos a continuación.

Los datos de geolocalización son de gran interés en diversos contextos. Se obtienen, por ejemplo, a través de móviles y distintas apps en ellos instalados. Como ejemplo, la Figura 1 izda muestra la traza

FIGURA 2
IZDA: IMAGEN ORIGINAL, CORRECTAMENTE CLASIFICADA COMO UN 2. DCHA: IMAGEN LIGERAMENTE PERTURBADA, INCORRECTAMENTE CLASIFICADA COMO UN 7



Fuente: Elaboración propia

durante un día de un usuario de una app de pago que empleamos en un proyecto de geomarketing. Basado en esos datos (a lo largo de varios días) es relativamente sencillo, con técnicas de filtrado y de análisis de conglomerados, encontrar el hogar y el lugar de trabajo (si es fijo) de los individuos y producir agrupaciones de clientes en función de los mismos, véase la Figura 1 dcha. Tal tipo de análisis tiene usos relevantes p.ej. en planificación urbana. Para obtener valor adicional debemos cruzar tales datos con información de otro origen para dar contenido semántico a los lugares que visita el individuo. Una posibilidad es cruzar las coordenadas obtenidas con bases de datos geolocalizadas, por ejemplo, asociadas a OpenStreetMap que incluyen, además, la tipología del lugar correspondiente (centro de estudios, restaurante, sinagoga,...) Con esa información podemos encontrar un perfil detallado de los hábitos espacio-temporales individuales, incluyendo su perfil como viajante, su religión, su edad, sus patrones de consumo de servicios, su nivel socio-económico,... (4) A partir de ahí, podemos segmentar los individuos con fines comerciales y predecir su tránsito temporal para enviarles publicidad en el momento adecuado.

En la descripción anterior se habrá apreciado las posibles violaciones a la privacidad: es relativamente sencillo controlar dónde nos ubicamos y predecir donde nos vamos a ubicar, así como predecir lo que haremos en tales ubicaciones. Estas capacidades se potencian si, además, se cruza la información anterior con el análisis de textos emitidos por el usuario como tuvimos que realizar en un proyecto de marketing en redes sociales. A partir de ellos, podemos inferir los rasgos de personalidad, así como su propensión a compra de ciertos productos. Esto nos permitiría modular los mensajes a enviar a los individuos para persuadirles mejor de ciertas opciones. Combinando ambos perfiles tenemos pues información de dónde y cuándo se va a ubicar una persona, qúe cosas le interesan y cómo debemos comunicarle esas cosas para incrementar su interés. Obviamente esta modelización es casi un Nirvana para el marketing comercial. Pero también lo es en el ámbito político y abre la puerta a escándalos como el de Facebook y Cambridge Analytica en relación al Brexit. El RGPD pone, sin embargo, fuertes restricciones sobre la posibilidad de perfilado de individuos.

Protección de la toma de decisiones

Los éxitos en la Economía del Dato están haciendo que se desplieguen de manera creciente sistemas basados en IA. Sin embargo, algunos de ellos son vulnerables a ataques maliciosos, lo que introduce riesgos obvios sobre, por ejemplo, sistemas de filtro de contenidos, ADSs o sistemas de defensa (Comiter, 2019). En consecuencia, estos deberían ser robustos en sus respuestas frente a tales ataques, si queremos confiar en operaciones basadas en sus salidas. Los algoritmos de última generación funcionan extraordinariamente bien frente a datos estándar, pero han demostrado ser vulnerables frente a ejemplos adversarios, instancias de datos dirigidas a engañarlos (Goodfellow *et al*., 2014). En tales contextos, los algoritmos deben diseñarse para tener en cuenta la posible presencia de adversarios y protegerlos frente a manipulaciones de datos. Como hipótesis fundamental, los sistemas basados en IA suponen típicamente el uso de datos independientes e idénticamente distribuidos, tanto para el entrenamiento como para las operaciones. Sin embargo, los aspectos de seguridad en el aprendizaje profundo, parte del campo emergente del aprendizaje automático adversario (AML), cuestionan dicha hipótesis, dada la presencia de adversarios dispuestos a intervenir en el problema para modificar los datos y obtener un beneficio. Como ejemplo motivador, los algoritmos de visión son el núcleo de varias tecnologías, como los ADS. Los ejemplos de ataques a tales algoritmos más sencillos y conocidos consisten en modificaciones de imágenes de manera que la alteración se

vuelve irrelevante para el ojo humano a efectos de reconocimiento, pero hace que un modelo entrenado en millones de imágenes clasifique erróneamente las atacadas, con consecuencias de seguridad potencialmente relevantes. Por ejemplo, con un modelo de red convolutiva relativamente sencillo, podemos predecir con precisión del 99% los dígitos escritos a mano en el conjunto de datos MNIST (LeCun *et al*., 1998). Sin embargo, si atacamos esos datos con el método rápido del signo del gradiente (Szegedy *et al*., 2013), la precisión se reduce hasta el 62%. La Fig. 2 proporciona un ejemplo de una imagen MNIST original y una atacada: a nuestros ojos ambas imágenes parecen un 2, pero el clasificador profundo identifica correctamente un 2 en el primer caso, mientras que sugiere un 7 tras el ataque.

El AML es un área difícil que evoluciona rápidamente y está conduciendo a una verdadera carrera de armamento en la que la comunidad alterna ciclos de proposición de ataques con ciclos de presentación de defensas frente a los mismos. Sin embargo, se basa en teoría de juegos estándar, con condiciones fuertes de conocimiento común no sostenibles. En Rios Insua *et al*. (2020b) se propone una metodología basada en la teoría de la decisión bayesiana para resolver problemas de AML, adoptando una perspectiva desde el análisis de riesgos adversarios que evita esas consideraciones. Rios Insua *et al*. (2020a) aplica este marco a la clasificación adversaria ilustrando la mayor robustez frente a ataques adversarios que se obtiene con tal aproximación.

Interpretabilidad de las decisiones automatizadas

Muchos de los algoritmos anteriormente descritos, por ejemplo los basados en redes profundas, tienen escasa capacidad explicativa, lo que puede motivar dudas éticas sobre su aplicabilidad para apoyar la toma de decisiones en aplicaciones críticas, problema que se acentúa si se detectan sesgos cuando tales decisiones discriminan a algún colectivo.

Los algoritmos de aprendizaje automático se han empleado con éxito para desarrollar IAs capaces de actuar de manera sobrehumana (5) en juegos conocidos como el ajedrez, el póquer o el go (y en otras actividades de mayor impacto social). La forma de estas IA, a menudo basadas en redes profundas, dificulta que un ser humano comprenda cómo el sistema computacional toma sus decisiones. Si bien descubrir estrategias sobrehumanas es un objetivo importante, es igualmente relevante comprender el razonamiento subyacente que explica por qué estas estrategias son superiores. Las IA de "caja negra", aunque resultan ser estrategas excepcionalmente buenas, nos dejan preguntándonos demasiadas veces ¿Cómo hizo eso la máquina?, cuestión especialmente importante en áreas sensibles como defensa, sanidad o finanzas, que requieren enfoques transparentes, responsables y comprensibles. En este punto, vale la pena mencionar cómo el RGPD podría requerir que los proveedores de IA proporcionen explicaciones sobre los resultados de la toma de decisiones automatizada basada en datos personales. (6) Esta falta de transparencia ha llevado a un interés creciente por una subárea del aprendizaje automático conocida como IA explicable (XAI) que, aunque puede tener varios significados dependiendo del contexto, debe cumplir dos requisitos: en primer lugar, como cualquier IA, debe ser capaz de tomar buenas decisiones o adoptar inferencias precisas; en segundo debe explicar fácilmente a los no expertos cómo llegó a sus conclusiones. Según esta definición, una IA que predice el clima utilizando un modelo matemático de dinámica atmosférica sería explicable; una basada en redes profundas no lo sería, típicamente.

Hay varios enfoques a este problema que se revisan a fondo en Burkart y Huber (2021). Una posibilidad importante es emplear modelos interpretables, fácilmente comprensibles para los humanos, como argumenta convincentemente Rudin (2019) que afirma que, en muchos contextos, tales modelos pueden funcionar casi tan bien como las redes neuronales profundas. Si bien el uso de modelos interpretables puede ser adecuado en algunos contextos, tiene el coste de su flexibilidad, precisión y usabilidad.

Otro problema interesante, y más fundamental, se refiere a aprender las reglas del juego. La mayoría de sistemas de IA típicamente llevan las reglas del juego preprogramadas y su entrenamiento consiste en aprender a escoger una estrategia ganadora del conjunto de movimientos factibles basado en el estado actual del juego. Así, el sistema tiene ventaja sobre un principiante que debe primero aprender las reglas antes de aprender una buena estrategia. La idea de que una máquina aprenda las reglas del juego ha motivado investigación en sistemas que aprenden observando cómo otros juegan, véase p.ej. Bjornsson (2012). En una contribución muy reciente (Aurentz *et al*., 2022) hemos presentado una IA interpretable, y su correspondiente algoritmo de aprendizaje automático, capaz de aprender en tiempo polinómico las reglas de un juego siempre que las relaciones entre el estado de un jugador y sus movimientos factibles pueda representarse mediante un conjunto de polinomios de Zhegalkin de grado bajo. Además, las reglas se almacenan de forma económica y producen una representación fácil de interpretar y transcribible a lenguaje natural.

ECONOMÍA DEL DATO. UNA REFLEXIÓN DESDE LOS ODS

A través de ejemplos aplicados hemos visto el potencial que tienen los métodos de la Inteligencia Artificial y del Big Data en la Economía del Dato para, por un lado, promover el bien social pero, por otro, poner en riesgo algunos valores comúnmente aceptados en la sociedad occidental. Podemos poner en perspectiva todos los proyectos descritos por medio de los denominados objetivos de desarrollo sostenible (ODS) de la Agenda 2030, (7) como se indica en

CUADRO 1
RELEVANCIA DE ALGUNOS ODS EN LOS PROYECTOS PRESENTADOS

ODS	Meta	Proyecto
3	Asegurar vidas saludables	2.1
9	Infraestructuras fiables	2.2
3	Reducir muertes por tráfico	2.3
9	Modernizar infraestructuras	2.3
11	Acceso sostenible a transporte	2.3
4	Asegurar educación inclusiva	2.4
16	Desarrollar instituciones auditables	3.1
16	Promover el estado de derecho	3.2
16	Crear instituciones transparentes	3.3

Fuente: Elaboración propia

el Cuadro 1, comprobándose el impacto positivo de los proyectos.

Como hemos indicado, por el momento en la Economía del Dato han predominado las ideas de negocio. Sin embargo, existe un enorme potencial en las aplicaciones en el ámbito social para beneficio de las administraciones y organizaciones no gubernamentales, como hemos ilustrado anteriormente. A pesar de ello, pocas decisiones gubernamentales se benefician aún del aprovechamiento sistemático de grandes masas de datos y técnicas avanzadas de modelización. Por comparación con las aplicaciones industriales, no es difícil vislumbrar las enormes aplicaciones que tendrían en problemas relativos al desarrollo racional de planes para infraestructuras; 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; o la identificación de barrios con servicios sociales inadecuados, entre otros muchos.

Surge entonces, el uso de la analítica para apoyar la toma de decisiones en la elaboración de políticas públicas, que denominamos Analítica para Políticas (Policy Analytics) (Daniell *et al.*, 2016). El ejemplo mencionado de AESA, y otras experiencias recientes del Instituto Nacional de Estadística en relación con sus estadísticas experimentales, muestran los enormes ahorros potenciales de los que nuestro país podría beneficiarse con una aplicación coherente y sistemática de las metodologías propuestas, que podrían planificarse de manera estratégica a partir de los mencionados ODS.

DISCUSIÓN

A través de ejemplos hemos mostrado el potencial de los sistemas basados en el aprendizaje automático y el razonamiento estadístico para fomentar un nuevo florecimiento de la sociedad. Pero también su potencial para afectar a valores tradicionalmente aceptados, al menos en la cultura europea. En función de las vías que adoptemos conformaremos nuestra sociedad en el futuro.

Concluimos recordando algunas de los principios que emanan de los ejemplos presentados y que son relevantes en el futuro desarrollo de la Economía (Pública) del Dato:

- Es importante incidir en el aspecto del valor que aportan tales datos. Acumular datos meramente puede ser inútil e ineficiente.
- La altísima carga de modelización requerida lleva típicamente a la necesidad de automatizar el tipo de tareas descrito.
- El preprocesamiento de datos es un proceso complejo y poco lucido, del que se habla poco, pero que resulta esencial para hacer descubrimientos interesantes y crear valor en la sociedad.
- En algunos dominios resulta esencial ser capaces de explicar los resultados de los sistemas basados en IA. Además, en algunos dominios, resulta suficiente emplear métodos interpretables.
- El tipo de desarrollos descritos está posibilitando nuevas tecnologías (como la de los vehículos autónomos).
- Igualmente, conlleva problemas morales que pueden resolverse a través de los métodos del análisis de riesgos.
- A partir de la huella digital de un individuo es posible inferir numerosas propiedades del mismo, lo que puede hacernos muy vulnerables.

Concluimos con algunas reflexiones finales:

- Es posible que los análisis de Big Data hayan sido de alguna manera sobrevendidos por las consultoras TIC. Por ejemplo, hemos podido leer expresiones del estilo *El diluvio de datos vuelve obsoleto el método científico*, (8) como si sólo necesitásemos recopilar grandes cantidades de datos y, a través de soluciones automatizadas, obtener algún tipo de modelo automatizado para tratar cualquier problema que podamos imaginar. Algunos avances recientes en química y biología, véase p.ej. Gallego *et al*. (2022), van en tal dirección, pero esa propuesta, sin duda, ignora algunos aspectos importantes de la ciencia. Por ejemplo, aunque los datos son importantes, debemos reconocer que, en muchos problemas, no habrá tantos. E incluso si los hubiere, aún existe una clara necesidad de incluir juicios de expertos y otras tecnologías analíticas en los procesos de toma de decisiones y de simulación de políticas para obtener aproximaciones más eficientes.

- Debemos insistir en que Big Data no se refiere sólo a tecnología (a Hadoop, Spark o similares) sino que requiere además metodologías científicas de la estadística y del aprendizaje automático, parte de lo que hoy llamamos ciencia de datos, y conocimientos sobre la materia en la que se hace un proyecto Big Data. (9)
- Desde el punto de vista tecnológico, debemos esperar, como no puede ser de otra manera, una evolución permanente en un fenómeno que apenas acaba de comenzar.
- Desde una perspectiva metodológica, debemos esperar también una importante evolución. Destacamos tres aspectos:
    - Un campo esencial sería el desarrollo de métodos escalables para inferencia bayesiana. Su status quo hace que prevalezcan de nuevo los métodos de máxima verosimilitud en este dominio, que tienden a ignorar la incertidumbre epistémica relacionada con el conocimiento, crucial para desarrollar una IA más segura y justa.
    - También resulta importante la integración coherente de estas metodologías en sistemas de ayuda a la toma de decisiones: el objetivo final de los modelos de inferencia y predicción debe ser la ayuda a la toma de decisiones y la Teoría de la Decisión (French y Rios Insua, 2000) facilita un marco normativo adecuado para tal integración.
    - Finalmente, se suele mencionar el aprendizaje por refuerzo como aproximación a una IA general; sin embargo, el énfasis debería ponerse en aprendizaje por refuerzo multiagente. Incidentalmente, aquí prevalece los conceptos de teoría de juegos pero conllevan condiciones de conocimiento común no sostenibles en muchos dominios, por lo que convendría desarrollar las aproximaciones basadas en el análisis de riesgos adversarios (Banks *et al*., 2015).
- Finalmente, desde el punto de vista ético, sería necesaria una mayor concienciación de la población respecto al valor de los datos y la regulación de los aspectos de privacidad. En particular, serían importantes campañas de comunicación en el corto plazo, y de educación en el medio plazo, que pongan de manifiesto la necesidad de disponer de datos y procesos de calidad reutilizables en un marco de transparencia.

Agradecimientos. Los proyectos concretos mencionados incluyen al AXA Research Fund (a través de la Cátedra AXA-ICMAT en Análisis de Riesgos Adversarios), la Agencia Estatal de Seguridad Áerea, la Comisión Europea (a través del proyecto Trustonomy con el código 815003), Aisoy Robotics, A3sec, la Fundación BBVA, el Real Instituto Elcano, Xeerpa, Quirónprevención, la National Science Foundation y la European Office for Aerospace Research and Development.

NOTAS

[1] El documento https://www.ine.es/normativa/leyes/cse/papel_estadistica_oficial.pdf del Consejo Superior de Estadística, proporciona información detallada sobre esta cuestión.
[2] https://www.un.org/es/about-us/universal-declaration-of-human-rights
[3] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[4] Curiosamente, no es fácil sin embargo inferir el género del individuo con este tipo de datos.
[5] En el sentido de que vencen al mejor jugador humano.
[6] Esto incluso conduciría a la prohibición de modelos opacos en ciertos dominios de aplicación.
[7] Véase https://unstats.un.org/sdgs/indicators/indicators-list/.
[8] En un artículo de Anderson en Wired en 2008.
[9] En los ejemplos de la sección 2 serían Cardiología, Seguridad Aérea, Ingeniería de Transportes y Robótica.

REFERENCIAS

Ashton, M. C., Lee, K., y de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review*, 18(2):139–152.

Aurentz, J., Navarro, A., y Rios Insua, D. (2022). Learning the rules of the game: An interpretable ai for learning how to play. *IEEE Transactions on Games*.

Banks, D. L., Rios, J., y Rios Insua, D. (2015). *Adversarial risk analysis*. CRC Press.

Bjornsson, Y. (2012). Learning rules of simplified boardgames by observing. In *Proc. ECAI 2012*, pages 175–180.

Brynjolfsson, E. y Kahin, B. (2002). *Understanding the digital economy: data, tools, and research*. MIT press.

Burkart, N. y Huber, M. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Burns, L. y Shulgan, C. (2019). *Autonomy: The Quest to Build the Driverless Car—And How It Will Reshape Our World*. ECCO.

Caballero, W., Naveiro, R., y Rios Insua, D. (2022). Modeling ethical and operational preferences in automated driving systems. *Decision Analysis*.

Castillo, E., Gutierrez, J. M., y Hadi, A. S. (2012). *Expert Systems and Probabilistic Networks*. Springer.

Claussmann, L., Revilloud, M., Gruyer, D., y Glaser, S. (2019). A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–23.

Comiter, M. (2019). *Attacking Artificial Intelligence*. Belfer Center Paper.

Cox, T. (2008). What's wrong with risk matrices. *Risk Analysis*, 28:497–512.

Daniell, K., Morton, A., y Rios Insua, D. (2016). Policy analysis and policy analytics. *Annals of Operations Research*, pages 1–13.

Elvira, V., Bernal, F., Hernandez-Coronado, P., Herraiz, E., Alfaro, C., Gómez, J., y Ríos Insua, D. (2020). Safer skies over spain. *INFORMS Journal Applied Analytics*, 50:21–36.

French, S. y Rios Insua, D. (2000). *Statistical Decision Theory*. Wiley.

Gallego, V., Naveiro, R., Roca, C., Campillo, N., y Rios Insua, D. (2022). Ai in drug development: a multidisciplinary perspective. *Molecular Diversity*.

Gardner, H. (2011). *Frames of mind: the theory of multiple intelligences*. Hachette UK.

Goodfellow, I. J., Shlens, J., y Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Jarvis Thomson, J. (1985). The trolley problem. *Yale Law Journal*, pages 1395–1415.

LeCun, Y., Cortes, C., y Burges, C. (1998). THE MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Liu, S. y Rios Insua, D. (2020). An affective decision-making model with applications to social robotics. *EURO J Decis Process*, 8:13–39.

Loewenstein, G. y Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, 619(642):3.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4):370.

McAllister, R., Gal, Y., Kendall, A., van der Wilk, M., Shah, A., Cipolla, R., y Weller, A. (2017). Concrete problems for autonomous vehicle safety: advantages of bayesian deep learning. In *Proc. 26th IJCAI*.

Nielsen, T. y Jensen, F. (2008). *Bayesian Networks and Decision Graphs*. Springer, New York.

Picard, R. W. (1997). Affective Computing. *Encyclopedia of Multimedia Technology and Networking, Second Edition*, (321):15–21.

Rios Insua, D., Caballero, W., y Naveiro, R. (2021a). Managing driving modes in automated driving systems. *arXiv: 2107.00280*.

Rios Insua, D., Camacho, J. M., Santos, A., y Lozano, A. (2021b). A predictive bayesian network model for cardiovascular diseases. Technical report, ICMAT.

Rios Insua, D. y Gómez-Ullate, D. (2019). *¿Qúe sabemos de? Big Data*. La Catarata.

Rios Insua, D., Naveiro, R., y Gallego, V. (2020a). Perspectives on adversarial classification. *Mathematics*, 8(11).

Rios Insua, D., Naveiro, R., Gallego, V., y Poulos, J. (2020b). Adversarial machine learning: Perspectives from adversarial risk analysis. *arXiv preprint arXiv:2003.03546*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Russell, J. A. y Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819.

Society of Automobile Engineers (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Technical report, SAE.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., y Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Rayner, M., y Townsend, N. (2017). European cardiovascular disease statistics. Technical report, European Heart Network.

Wu, B., Iandola, F., Jin, P. H., y Keutzer, K. (2017). Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 129–137.