IUS ET SCIENTIA
2022 • Vol. 8 • Nº 1• ISSN 2444-8478
https://editorial.us.es/es/revistas/ius-et-scientia
https://dx.doi.org/10.12795/IETSCIENTIA • © Editorial Universidad de Sevilla 2022

La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo de una noción no normativa de salud

THE CHIMERA OF ALGORITHMIC OBJECTIVITY: DIFFICULTIES OF MACHINE LEARNING IN THE DEVELOPMENT OF A NON-NORMATIVE NOTION OF HEALTH

Ariel Guersenzvaig Elisava
Universitat de Vic-Universitat Central de Catalunya
aguersenzvaig@elisava.net 0000-0002-6346-1512

David Casacuberta
Universitat Autónoma de Barcelona, Barcelona, España.
david.casacuberta@gmail.com 0000-0001-7119-9342

Recibido: 11 de diciembre de 2021 | Aceptado: 20 de mayo de 2022

RESUMEN

Este ensayo explora si el aprendizaje automático, una subdisciplina de la inteligencia artificial, puede contribuir a desarrollar un acercamiento más objetivo al desarrollo y formulación de conceptos y descripciones, tomando como ejemplo el caso de la definición de salud. Para ello se aborda la teoría naturalista de la salud propuesta por Christopher Boorse y se la contrasta con una serie de posibilidades y problemas que pueden surgir al aplicar el aprendizaje automático a la formulación junto a esta teoría. En base al análisis se concluye que tanto el aprendizaje automático (tanto supervisado como no supervisado) arrastran elementos de normatividad y subjetividad que hacen inviable el desarrollo de conceptos y descripciones de manera neutra y objetiva. Esto no implica que el aprendizaje automático quede invalidado para el análisis evaluativo de la salud, sino que resalta y explicita los elementos subjetivos presentes en él.

PALABRAS CLAVE

Aprendizaje automático
Salud
Objetividad
Normatividad

ABSTRACT

This essay explores whether machine learning, a sub-discipline of artificial intelligence, can contribute to developing a more objective approach to the development and formulation of concepts and descriptions. Taking as an example the case of the definition of health proposed by Christopher Boorse, the paper discusses and contrasts a series of possibilities and problems that

KEYWORDS

Machine learning
Health
Objectivity
Normativity

may arise when applying machine learning to solving some of the problems encountered by this theory. Based on the analysis, the paper concludes that machine learning (both supervised and unsupervised) entail elements of normativity and subjectivity that make it unfeasible to develop concepts and descriptions in a neutral and objective manner as the theory requires. This does not imply that machine learning is invalidated for the evaluative analysis of health, but rather highlights and makes explicit the subjective elements present in it.

En este ensayo exploramos si el aprendizaje automático, una subdisciplina de la inteligencia artificial, puede contribuir a desarrollar un acercamiento más objetivo al desarrollo y formulación de conceptos y descripciones, tomando como ejemplo el caso de la salud. En una primera sección trataremos la cuestión de qué implica hablar de objetividad de datos y algoritmos en el mundo de la inteligencia artificial. Seguidamente abordaremos en detalle un buen candidato para esta función, la teoría naturalista de la salud propuesta por Christopher Boorse. Para ello presentaremos primero la teoría en cuestión y comentaremos una de las principales críticas planteadas. Después de un breve desvío para introducir muy brevemente los conceptos principales del aprendizaje automático, retomaremos el mello de la cuestión considerando una serie de posibilidades y problemas que pueden surgir al aplicar aprendizaje automático a la formulación de una teoría naturalista de la salud.

I. Objetividad, datos e inteligencia artificial

Es común la idea de que los resultados obtenidos por un algoritmo de inteligencia artificial, al contrario de una respuesta producida por un humano, son, en sí mismos objetivos, no distorsionados por ningún tipo de sesgos. Los argumentos para sostener esta posición son variados.

Por un lado tenemos argumentaciones originadas y compartidas por el público en general, creando una especie de imaginario colectivo de qué es un robot, un software de inteligencia artificial, alimentadas por las películas y novelas de ciencia-ficción. Las máquinas no tienen emociones, por lo tanto, son cien por cien racionales y no se dejan engañar por ira, depresión o alegría excesiva. En la misma línea se apuntan a ideas como "las máquinas nunca se cansan", "las máquinas no olvidan las instrucciones", "las máquinas no cometen errores y siguen su programación al milímetro", etc. etc.

Por otro lado, tenemos argumentos que vienen sobre todo del mundo de la ingeniería informática que defienden que un sistema de inteligencia artificial, si se desarrolla de forma correcta, será –de manera natural– objetivo y no sesgado. Si el programa acaba teniendo sesgos será por qué los han introducido –de forma deliberada o inconsciente– las personas que han desarrollado el *software*.

Esta visión de las ciencias de la computación se construye básicamente a partir de estas dos premisas:

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

1. Los datos en sí mismos son objetivos. Es la interpretación de la teoría o modelo construida a partir de ellos lo que puede ser subjetivo o sesgado
2. Las matemáticas son un sistema puramente formal. No tiene contenido. Por lo tanto no pueden ser sesgadas por definición.

¿Son ciertas estas premisas? Realmente un algoritmo es, casi por definición, objetivo? Vamos a examinar estos dos modelos y veremos como ninguno de los dos nos garantiza una objetividad de salida.

La visión popular de qué es la inteligencia artificial y por qué es objetiva es la más fácil de refutar. No hay que subestimar el impacto e influencia de la ciencia-ficción en este imaginario colectivo de la objetividad robótica. Tenemos que pensar que las novelas y películas de ficción científica nos hablan del futuro. Sobre el papel –o el celuloide, o el byte– es así, pero si analizamos con cierto detalle este género descubriremos que este futuro es básicamente una metáfora para hablar del presente. Cuando George Orwell buscaba un título para su distopía sobre un estado futuro totalitario, decidió intercambiar las últimas cifras del año en que escribió la novela, 1948, y así nació 1984. Orwell no quería tanto advertirnos de un lejano futuro distópico en el que somos sistemáticamente engañados, controlados y vigilados como describir las prácticas totalitarias, bien actuales entonces, del régimen de Stalin y explorar de qué forma las tecnologías futuras podrían exacerbar esas tendencias.

De la misma forma, el imaginario objetivo de la inteligencia artificial como objetiva, racional y sin sesgos no es tanto una teorización sobre cómo será el futuro sino una guía bien presente de qué esperamos aquí y ahora de la inteligencia artificial.

En este imaginario colectivo se combinan deseos y temores. Por un lado está la construcción de la inteligencia artificial como una forma de traer más racionalidad al mundo, pero por otro está el temor de que esas máquinas nos conviertan en obsoletos y decidan rebelarse contra nosotros.

Este temor a máquinas que nos superan y finalmente deciden eliminarnos se construye paradójicamente con los mismos conceptos de los que surge la utopía de la máquina racional perfecta que no tiene sesgos. Esa racionalidad implacable rápidamente origina visiones de eficiencia despiadada, de carencia total de empatía, que fácilmente nos hace visualizar un apocalipsis robótico. De hecho si rascamos un poco en las propuestas de Bostrom (2014), sobre la "superinteligencia" y sus peligros, veremos que la idea de fondo también es la misma: una inteligencia artificial general (es decir, una que puede cambiar de tema de exploración o reflexión y ser igualmente eficaz digamos jugando al ajedrez, conduciendo un automóvil o escribiendo haikus) sería muy superior a la humana y claramente podría decidir que los humanos sobramos de la escena[1].

1. No deja de ser curioso cómo esta idea ha ido evolucionando y cada vez se acepta de forma más natural. En la película *2001 Una Odisea del Espacio* Clarke y Kubrick tuvieron que inventarse un problema de conflicto de órdenes en la programación del ordenador HAL 9000 que le causa finalmente una especie de paranoia por la que intenta acabar con toda la tripulación. En cambio, ficciones posteriores como la serie *Terminator* postulan como algo lógico y natural que una vez que la red *Skynet* –por error humano– consigue ser autoconsciente, su primera decisión es provocar una guerra nuclear que

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Así pues, podemos dejar de lado esta caracterización del público general de las máquinas como naturalmente objetivas ya que no se fundamentan en ningún tipo de modelo teórico sino simplemente en el imaginario presentado por la ciencia-ficción. Los otros elementos que surgen del sentido común (las máquinas no tienen emociones, no se cansan, no necesitan dormir, etc.). No son ni necesarios ni suficientes para establecer la objetividad de las decisiones. Si bien es cierto que una doctora puede cometer un error al diagnosticar un paciente al estar irritada por haber dormido poco, no es menos cierto que la inmensa mayoría de las veces esa doctora es capaz de ver más allá de su cansancio y establecer el diagnóstico correcto. Y de la misma forma, que yo me levante fresco como una rosa después de una buena siesta no garantiza que todas las decisiones que tome a partir de ahí sean correctas, racionales y sin sesgos.

La visión de la objetividad que surge de la ingeniería es mucho más elaborada y consistente, pero finalmente, también es incorrecta.

Empecemos por la idea de que los datos en sí mismos son objetivos y es la teoría posterior la que puede *subjetivizarlos*. Esta es una idea que cada vez tiene más adeptos, y tiene su relato perfecto en Anderson (2008) donde se habla directamente de un "fin de la teoría" en el que el análisis computacional mediante algoritmos de aprendizaje automático acabará haciendo obsoleto el propio método científico (Anderson 2008). Para Anderson, los datos son de salida, objetivos, y una teoría enmaraña con conceptos, conexiones causales y otras malas hierbas filosóficas lo que es una conexión directa entre el dato y el efecto posterior. Un algoritmo de aprendizaje automático cribará los millones de datos disponibles y establecerá una serie de correlaciones útiles para hacer futuras predicciones. Dejamos así la teoría y nos instalamos en un *data deluge*, de predicciones fiables y ajustadas, renunciando a la comprensión humana a cambio de la fiabilidad, objetividad y precisión robótica.

Desgraciadamente, esta visión *tecnoutópica* de Anderson –y una parte importante del mundo de las ciencias de la computación - no se sostiene por varias razones.

En primer lugar, cualquier base de datos está, en su naturaleza, sesgada. No podemos compilar todos los datos del universo, con lo que tendremos que hacer una selección, un muestreo de datos, e inevitablemente, en esa selección tendemos a seleccionar unos elementos sobre otros. Tenemos actualmente decenas de ejemplos así, algoritmos supuestamente objetivos que después cometen graves sesgos pues las bases de datos de partida no representaban de forma equitativa las poblaciones, como el algoritmo de aprendizaje automático que etiquetaba manchas en la piel para establecer posibles melanomas y que funcionaba muy bien con personas de piel blanca pero cometía errores con personas de piel más oscura porque los individuos no caucásicos no estaban bien representados en esa base de datos (Adamson y Smith, 2018)

En segundo lugar, la visión del dato como algo objetivo se basa en una imagen muy ingenua y desactualizada de lo que es la ciencia. Desde la filosofía de la ciencia con tra-

eliminará prácticamente a toda la humanidad. Como si librarse de los humanos una vez un ser superior consiguiera la consciencia fuera un paso obvio. Algo así como librarse de una plaga de hormigas que ha aparecido en casa. No tenemos demasiado buena opinión de nosotros mismos…

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

bajos como van Fraassen (1980) o Feyerabend (1993) –e incluso el mismo Popper– está aceptado que cualquier observación está siempre cargada de cierta teoría que nos dice lo que hay que ver. La mayoría de "datos" en una base de datos se fundamentan en realidad conceptos teóricos. Un melanoma está muy lejos de ser un "dato crudo" reducible a datos inmediatos de la consciencia a la *Aufbau* de Carnap (1998). Si hacemos caso a van Fraassen (1980) incluso una partícula subatómica como el electrón es en realidad un concepto teórico, que no se sigue necesariamente de las observaciones establecidas en un acelerador de partículas.

Y, finalmente, aún suponiendo que los datos con los que alimentamos a un supercomputador que predice el tiempo que hará son objetivos y no constructos teóricos, cualquier esperanza de objetividad en los datos se desvanece en las ciencias humanas. Si pensamos en la mayoría de categorías con las que clasificamos el comportamiento humano, veremos que la mayoría de ellas están cargadas de cierto elemento valorativo: sociabilidad, empatía, discapacidad, sociopatía, activo, pasivo, conservador, demócrata, totalitario, etc. son todos términos que van asociados a una escala de valores y a una concepción concreta del mundo humano. Su inclusión en cualquier base de datos nos lleva, de salida a una determinada y específica interpretación de la realidad, una concepción inevitablemente sesgada de qué es la humanidad.

Recordemos así la segunda premisa de la algoritmización de la ciencia como una disciplina objetiva. La imposibilidad de los sesgos en las matemáticas. Para buena parte del mundo de las ciencias de la computación y la ingeniería la idea se cae por su propio peso. Que las matemáticas puedan estar sesgadas les parece una insensatez, quizás lo encuentren hasta irritante. ¿Cómo va a ser sexista el teorema de Pitágoras? ¿De qué manera podría convertirse en racista la tabla de multiplicar? Piensan que una declaración así solo puede surgir de los desvaríos irracionalistas de un radical filósofo posmoderno.

En una reciente pieza de opinión para *The Spectator,* el investigador en ciencias de computación e inteligencia artificial Pedro Domingos (2021) defendía que por su propia naturaleza matemática, un algoritmo necesariamente está libre de sesgos:

> Los algoritmos de aprendizaje automático son simplemente complejas fórmulas matemáticas que no saben nada sobre raza, género o estatus socioeconómico. Son tan racistas como lo podría ser la fórmula y=ax+b.

¿Por qué nos resulta tan difícil aceptar que un algoritmo puede ser racista o sexista? Pensamos que esa creencia es resultado de una serie de confusiones conceptuales sobre lo que es una estructura formal como un algoritmo.

Domingos confunde aquí formalismo con objetividad. Que una estructura sea matemática, es decir puramente formal, indica que está vacía en sí misma de contenido asociado al mundo exterior. Pero si la fórmula se va a usar para resolver un problema en el mundo exterior, entonces las estructuras formales del algoritmo codifican aspectos del mundo exterior. Si la forma de codificación se basa en un sesgo racista o sexista, el resultado del algoritmo será racista o sexista, por muy matemático que sea el cálculo.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Un ejemplo sencillo para que nos entendamos.

Imaginemos un algoritmo que desarrolla de forma automática la contabilidad de una empresa, y entre otras cosas, procesa las nóminas de todos los empleados En esa base de datos el género de los empleados está codificado de manera binaria 0 es igual a hombre y 1 es igual a mujer. El programa incluye también una variable "sueldoMedio" que almacena lo que ha de cobrar un trabajador de salida.

Imaginemos ahora que el algoritmo incluye la siguiente línea, que aquí transcribimos en pseudocódigo:

Si trabajador.sexo = 1 restar 10% a sueldo medio

Se trata, como bien dice Pedro Domingos de una fórmula matemática que pide que a una cantidad X le restemos el diez por ciento, matemáticamente:

Y=X-X*10/100

Pero recordemos que 1 significa "mujer" de manera que cuando este algoritmo se aplique y calcule las nóminas, todas las mujeres de la empresa cobrarán un diez por ciento menos que los hombres. Es bastante absurdo defender que este algoritmo es "objetivo" porque se basa en "matemáticas".

Las bases de datos nunca son objetivas en sentido estricto. Están organizadas a partir de una representación determinada de la realidad. Hay personas que deciden qué características son relevantes y cuáles no lo son, qué se incluye en ellas y qué se deja fuera, así como también qué puede deducirse y cómo actuar en función de esos datos que se han recopilado.

Los sesgos a veces surgen directamente de las mentes sesgadas de las personas que toman decisiones. Si los ejecutivos de una empresa son sexistas y creen que las mujeres no están preparadas para acceder a puestos directivos, y usamos esas decisiones sesgadas para desarrollar un algoritmo de aprendizaje automático, el algoritmo inevitablemente tomará esas mismas decisiones sesgadas que esos ejecutivos machistas. Domingos tiene razón al afirmar que un algoritmo no sabe nada sobre raza, sexo o estatus económico. Pero de ahí no se sigue que los algoritmos sean objetivos. Lo único que se sigue es que copiarán a la perfección todos los sesgos que las bases de datos hayan capturado.

La respuesta habitual desde el mundo de la computación y la ingeniería es que ello es así, pero que los responsables son los humanos, no las máquinas. Y son optimistas argumentando que la inclusión de algoritmos de aprendizaje automático puede ayudar a convertir las ciencias humanas en verdaderas ciencias, al conseguir datos realmente objetivos que son procesados de forma infalible por fórmulas matemáticas. Un estudio matemático ayudaría claramente a descubrir sesgos conscientes de las personas que han desarrollado el algoritmo, como mi ejemplo de la línea de código que resta un 10% del sueldo medio por ser mujer. Pero podrían ir más allá, podrían descubrir sesgos inconscientes en la forma en que categorizamos que es una discapacidad perceptiva, un problema mental o un traumatismo y ayudarnos a tener unas ciencias más objetivas y una aplicación de sus resultados más justa.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Esta tendencia es especialmente relevante en el mundo de la salud. La biomedicina es un espacio muy abierto a los sesgos y las valoraciones y, por otro lado, tiene un impacto muy claro en el bienestar humano, por lo que conseguir una atención sanitaria más objetiva y menos sesgada es un objetivo central a conseguir según vamos incluyendo más algoritmos de inteligencia artificial en nuestras.

La pregunta que queremos explorar es esta: ¿Realmente los algoritmos de aprendizaje automático pueden ayudarnos a desarrollar una medicina no solo más acertada en sus diagnósticos, sino que además será más justa? Para establecer la verosimilitud de un acercamiento así, en este artículo nos centraremos en el concepto sobre el que inevitablemente todo el sistema ha de construirse: la idea de salud. "Salud" es un término escurridizo, muy difícil de definir, y que de salida es claramente valorativo. Todos queremos gozar de buena salud y evitar la enfermedad. "Salud" además de forma amplia incluye elementos complejos como el concepto de "salud mental" que implica una comprensión clara de cómo es una sociedad humana deseable y bien armonizada y qué comportamientos individuales la cuestionan.

¿Es posible ofrecer una definición objetiva de salud que luego usemos para generar algoritmos de aprendizaje automático más exactos y justos? Para establecer esa posibilidad necesitamos primero reflexionar qué entendemos por salud.

II. Teorías de la salud

Desde un punto de vista teórico, no existe consenso acerca de la noción de "salud", sino que encontramos diferentes aproximaciones y modelos teóricos al problema. En la gran mayoría de la literatura, al menos en la de enfoque analítico, se establecen dos aproximaciones antagónicas para definir la noción de "salud" (ver e.g. Kovács, 1998; Murphy, 2015). En uno de estos extremos encontramos la perspectiva conocida como *naturalista*, también a veces denominada *no normativista* u *objetivista*. En el otro encontramos perspectivas *normativistas*, también llamadas *subjetivistas* o *constructivistas*.

Las perspectivas naturalistas buscan ofrecer definiciones de salud que sean "neutras" en el sentido de estar libres de valores y ser objetivas. Los autores con esta perspectiva intentan ofrecer definiciones de salud en una manera similar al tipo de definiciones que encontramos en las ciencias naturales. Buena parte de estos modelos comparten la visión de los datos como entes de salida objetivos, que solo se *subjetivizan* cuando entran a formar parte de teorías que interpretamos, como hemos explicado en la sección primera. Decir que algo es "saludable" o está "sano" desde una posición naturalista es hacer una descripción neutra de una realidad.

La teoría naturalista más conocida y frecuentemente debatida es la Teoría Bioestadística de la Salud (*Bio-Statistical Theory*) desarrollada y revisada por Christopher Boorse a lo largo de cuatro décadas (Boorse, 1975, 1977, 1997, 2014). Esta teoría descansa en un entendimiento no normativo (i.e. neutro, objetivo) de la función biológica y en una noción estadística del concepto de "normalidad". En la tercera sección revisaremos esta teoría con mayor detenimiento.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Antes de pasar a debatir el tema vale la pena contraponer y comentar, aunque sea muy brevemente, la perspectiva opuesta para clarificar sus diferencias. Así, en oposición a las posturas objetivistas, las posturas normativistas argumentan que cualquier definición de salud irremediablemente involucra toda clase de normas socio-culturales y valores subjetivos, negando así la posibilidad que abríamos en la sección anterior de crear una biomedicina más objetiva a través de la inclusión de datos objetivos y algoritmos de inteligencia artificial. De esta manera, decir que algo está "sano" no es hacer una mera descripción de un hecho natural, tal como argumentan los naturalistas, sino que es realizar una evaluación normativa. Dentro de las corrientes normativistas encontramos una versión *fuerte* ("*strong*") que sostiene que los juicios acerca de la salud son *exclusivamente* valorativos y subjetivos, y una versión *débil* ("*weak*") que concede la existencia de elementos descriptivos objetivos a la vez que mantiene el carácter ineludiblemente subjetivo a la hora de determinar los procesos o estados a valorar (Boorse, 1975, p. 51). Para los normativistas, la salud no es, entonces, una realidad objetiva a la manera que puede serlo la composición mineral de una roca; en esta visión no hay un conjunto natural y objetivamente definible de funciones o disfunciones en relación a la salud.

1. Boorse y las nociones bioestadísticas de "salud" y "enfermedad"

Tal como avanzamos, Boorse intenta ofrecer una visión completamente objetiva y neutra de la cuestión; aquí, "salud", y también el término antagónico de "enfermedad", no son más que estados biológicos, objetivos. En este sentido, decir que un organismo está sano es hacer una *descripción* de un hecho natural, no es hacer una valoración acerca del mismo (en términos de bueno o malo, deseable o indeseable, etc.).

Para Boorse (1977, p. 542), "salud" es el funcionamiento *normal*, y "enfermedad" es "un estado interno que deprime la habilidad funcional por debajo del nivel típico de la especie" (Boorse, 1977, p. 542). En este sentido, evaluamos la salud de un organismo en relación a la especie a la que pertenece. Por ejemplo, así podemos determinar si una persona está sana o enferma comparándola con el funcionamiento normal de la especie *homo sapiens*.

Es con este enfoque que Boorse busca erradicar todo rastro de subjetividad de las nociones de salud y enfermedad. La idea central es que una condición patológica o enfermedad es un estado por debajo de la normalidad funcional estadística ("statistically species subnormal biological part-function") (Boorse, 1997, p. 4). Vale la pena insistir en que la "normalidad" debe ser entendida en sentido estadístico y la "función" en un sentido biológico. La salud es, entonces, la habilidad funcional normal. Para Boorse (1975, p. 57), "lo normal es lo natural", un organismo está sano cuando su funcionamiento normal es acorde a su diseño natural. En otras palabras, la salud es la idoneidad para desempeñar las funciones normales de un organismo con eficiencia estadísticamente normal en condiciones típicas (Boorse, 2014, p. 684). La función normal permite a un organismo asegurar su supervivencia y reproducción (Boorse, 1977, p. 555; 2014, p. 684).

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Además de la normalidad estadística relativa a la especie, para determinar si un organismo está sano o enfermo, Boorse introduce la noción de "clase de referencia" ("*reference class*"), "una clase natural de organismos de diseño funcional uniforme; específicamente un grupo etario de un sexo de una especie" (Boorse, 1977, p. 555). A modo de ejemplo, según la teoría bioestadística, para determinar si una persona tiene niveles normales de testosterona, debemos primero determinar su sexo y edad y usar estos datos como clase de referencia para comparar, ya que existe una importante variación estadística en los valores de testosterona entre hombres y mujeres de diferentes edades. Dado que para Boorse el diseño de la especie depende del sexo, la edad y, en algunos casos, de la raza, las abstracciones estadísticas deben hacerse en base a clases de referencia más pequeñas que las especies (Boorse, 1977, p. 558). Por ejemplo: "una mujer blanca de 35 años".

Mediante el uso de la noción estadística de funcionamiento típico de la especie, que *prima facie*, al ser una medida empírica, es una referencia descriptiva no normativa, Boorse parece haber dado suficientes elementos para evaluar de manera neutra y objetiva si alguien está sano o enfermo. Según Boorse, para realizar esta evaluación no se requieren juicios de valor y evaluar no implica necesariamente hacer una valoración.

III. Normatividad y clases de referencia

No es el objetivo de este ensayo presentar y comentar las múltiples objeciones que se han planteado a la teoría bioestadística de Boorse a lo largo de los años. En su lugar, nos concentraremos particularmente en una crítica a la teoría planteada por Elselijn Kingma (2007, 2014) en torno a los elementos de normatividad que se cuelan en la definición de las clases de referencia. Los lectores interesados en otras críticas pueden consultar algunos de los diferentes autores que han recopilado gran parte de estas objeciones (Ereshefsky, 2009, pp. 222-223; Gammelgaard, 2000, pp. 110-113; Murphy, 2015) y también las defensas ofrecidas por el propio Boorse (1997, 2014).

La objeción relacionada con las clases de referencia planteada por Kingma nos servirá como trampolín para explorar si el *aprendizaje automático* podría servir para resolver algunos de los desafíos que surgen a partir de ella. Antes de explicar de qué manera el aprendizaje automático podría contribuir a la teoría bioestadística de Boorse, es necesario primero comentar la objeción planteada.

Boorse muestra convincentemente que las comparaciones a nivel de especie resultan poco operativas y por ello introduce la idea de clase de referencia, que hemos examinado más arriba. Resulta evidente que para evaluar la normalidad de un estado biológico y clasificarlo como sano o enfermo, es necesario no solo comparar a nivel de especie, sino que se requiere también de una medida de referencia operativa para comparar organismos particulares. Si quisiéramos establecer la salud (o la normalidad) del corazón de un neonato deberíamos compararlo con otros neonatos y no con otros adultos, ya que, *normalmente*, el corazón de un neonato late mucho más deprisa que el de un adulto. Resultaría poco operativo comparar el latido de un neonato con el de un

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

adulto porque prácticamente la totalidad de los neonatos tendría un corazón que late *demasiado rápido* y sería considerado enfermo al no cumplir con el criterio de normalidad estadística de la especie.

Según Kingma (2007, p. 128), la teoría bioestadística es una descripción adecuada de salud únicamente si se eligen clases de referencia del tipo correcto. Que una persona sea considerada saludable no dependerá de la normalidad en relación a *cualquier* clase de referencia sino únicamente de la normalidad en relación a una clase de referencia "apropiada" (un bebé recién nacido debería ser evaluado en relación a la clase de referencia "neonatos" y no "adultos").

Kingma apunta que si eligiéramos diferentes clases de referencia que "sexo", "edad" o "raza", obtendríamos concepciones de salud muy distintas. Por ejemplo, si usáramos "fumadores habituales" como clase de referencia, entonces el cáncer y las EPOC serían consideradas estadísticamente "normales" y por tanto saludables. Es aquí donde cabe preguntarse, con Kingma, ¿por qué sería legítimo considerar "sexo", "edad" y "raza", y no otros criterios?

La respuesta que ofrece Kingma es que Boorse no puede justificar su elección de clases de referencia apropiadas sin involucrar juicios de valor y concepciones previas acerca de qué es saludable o no. Rechazamos "fumadores habituales" como clase de referencia porque entra en conflicto con nuestras intuiciones acerca de lo saludable. Naturalmente, también el propio Boorse rechazaría "fumadores habituales" como clase de referencia, pero este hipotético rechazo no haría sino enfatizar un punto débil de su propia teoría bioestadística. Si la teoría de Salud de Boorse busca ser neutra y objetiva, debería ser capaz de ofrecer una explicación sin juicios de valor acerca de qué criterios o tipos de criterios constituyen una clase de referencia apropiada y cuáles no.

La teoría bioestadística tiene muchas virtudes y es razonable admitir que una vez determinadas las clases de referencia es posible realizar análisis de salud y enfermedad libres de valores (Kingma, 2007, p. 132), es decir neutros y objetivos en el sentido operativo en el que estos términos son utilizados en las ciencias naturales. Sin embargo, dado que la elección de las clases de referencia no está libre de juicios de valor, si los argumentos de Kingma resultan convincentes, la teoría bioestadística de Boorse es inviable como una teoría naturalista en sentido estricto ya que admite elementos normativos.

En una línea similar, la filósofa Ruth Millikan (1984, pp.17-37) argumenta que cualquier clase biológica es fruto de un proceso de selección natural, de manera que para establecer cuál es la función adecuada de un mecanismo o proceso biológico no es suficiente con establecer estadísticamente cuál es la forma más común en que se presenta, sino que necesitamos establecer la historia evolutiva que ha llevado a ese organismo o proceso biológico a comportarse de cierta manera. Incluir esa historia evolutiva implica introducir así valoraciones sobre la forma más correcta en que el organismo o proceso se adaptaron a un entorno concreto.

Para ilustrar esta diferencia, Millikan (1984, p. 29) pone el ejemplo de los espermatozoides. Si los observamos de forma objetiva, estadística, veremos que el 99% de los espermatozoides no consiguen alcanzar el óvulo y simplemente desaparecen al cabo de un tiempo. Pero está claro que la función biológica del espermatozoide es fecundar

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

al óvulo y que los datos estadísticos, son por tanto, totalmente irrelevantes para establecer cuál es su verdadera función biológica.

IV. Es posible "naturalizar" y objetivar nuestra concepción de salud con los algoritmos de aprendizaje automático?

Hemos dicho que para aceptar la teoría bioestadística de Boorse como teoría naturalista, no alcanza con aseverar que "sexo", "raza" y "edad" son (las) clases de referencia apropiadas. En otras palabras. si la teoría bioestadística requiere de clases de referencia es preciso determinarlas y justificarlas de manera naturalista, es decir con una base empírica neutra y objetiva que prescinda de juicios de valor.

¿Sería posible mediante inteligencia artificial determinar, prescindiendo de elementos normativos, qué clases de referencia son apropiadas y cuáles no lo son?

La idea de utilizar inteligencia artificial para fines epistémicos no es particularmente nueva. En el anteriormente mencionado influyente texto de Chris Anderson "The End of Theory" (Anderson 2008), se afirma que los enormes volúmenes de datos y las herramientas para tratarlos ofrecen una nueva manera de entender el mundo en base a la correlación estadística entre datos. Esta correlación hace que las explicaciones basadas en el fenómeno de causación (es decir los modelos y teorías) devengan innecesarias para el avance científico (Anderson, 2008). Lo que propone Anderson no es utilizar inteligencia artificial para apoyar computacionalmente el descubrimiento científico, sino que sea la propia aplicación la que resuelva el problema, o directamente lo disuelva, haciendo así innecesaria la existencia de teorías que nos expliquen la realidad. Para ilustrar este tipo de aplicación de la inteligencia artificial, podemos citar el caso reciente de AlphaFold. Se trata de un sistema de inteligencia artificial que ha sido capaz de predecir la estructura en 3D de una proteína con altísima precisión, y resolver así uno de los grandes desafíos de la biología (Heaven, 2020).

En el resto de este ensayo no abordaremos de manera general las diferentes y profundas cuestiones epistemológicas que se pueden plantear acerca de la interacción entre inteligencia artificial y el descubrimiento científico (véase e.g. Casacuberta y Vallverdú, 2014). El objetivo aquí es mucho más modesto y se trata de explorar si mediante la inteligencia artificial sería posible determinar clases de referencia sin la intervención de elementos de subjetividad normativa. Si esto fuera posible, la teoría de Boorse superaría la importante objeción planteada por Kingma y estaría más cerca de ser una teoría naturalista de la salud más robusta.

4.1. ¿Qué es el aprendizaje automático?

Dado que el campo de la inteligencia artificial es extenso, nos concentraremos en la subdisciplina de la IA más popular en la actualidad: el "aprendizaje automático" (*machine learning*). Esta sección no busca ser exhaustiva sino únicamente ofrecer una breve introducción a la cuestión sin elementos técnicos.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

El aprendizaje automático, que estuvo originalmente vinculado a teorías cognitivistas y enfoques simbólicos, actualmente está basado en representaciones de conocimiento obtenidas mediante técnicas matemáticas y teoría estadística aplicadas en conjunción con el procesamiento computacional de enormes volúmenes de datos presentes en bases de datos (*big data*). Aunque existen otras técnicas de aprendizaje automático, el "aprendizaje profundo" (*deep learning*) es la más popular en la actualidad, en ella frecuentemente se utilizan "redes neuronales" (*neural networks*). Dada su conexión con las matemáticas y con el procesamiento algorítmico de datos estadísticos, el aprendizaje automático se presenta, entonces, como un candidato ideal para obtener clases de referencia neutras y objetivas.

Se suelen distinguir dos grandes tipos de aprendizaje automático: "aprendizaje supervisado" (*supervised learning*) y "no supervisado" (*unsupervised learning*) (Boden, 2016, pp. 48-49). En el aprendizaje *supervisado*, los programadores de un sistema lo entrenan definiendo una serie de resultados de salida esperados para una gama de datos de entrada, que son etiquetados (*labeled*) por el equipo de desarrollo. Una vez entrenado un modelo, el sistema es capaz de asignar una etiqueta de salida a un nuevo valor. Los usuarios o programadores del sistema pueden seguir entrenando el modelo con el propio uso, especificando al sistema si la etiqueta asignada es correcta. Por ejemplo, una *app* de reconocimiento de pájaros puede ser entrenada con enormes cantidades de fotos de pájaros etiquetadas con sus respectivos nombres genéricos (colibrí, ruiseñor, etc.). Una vez entrenado en modelo predictor, la *app* será capaz de asignar una etiqueta a una nueva imagen capturada por la app. Si la *app* ofrece una respuesta incorrecta (por ejemplo, etiquetando como "ruiseñor" a una "gaviota"), el usuario puede alertar de este error al sistema, que se retroalimentará a partir del error y ajustará sus hipótesis predictivas futuras.

El aprendizaje *no supervisado* no está basado en que los programadores etiqueten los datos de entrenamiento o especifiquen resultados concretos esperados, sino que la idea es que el propio sistema de manera autónoma detecte y reconozca patrones existentes en los datos (correlaciones estadísticas). Una técnica frecuente es el agrupamiento (*clustering*) en la cual se generan un conjunto de agrupamientos mediante la minimización o maximización de algún criterio de optimización. Por ejemplo, mediante el aprendizaje no supervisado, a partir de bases de datos de *marketing*, una empresa puede realizar una clasificación de sus clientes en distintos segmentos. Aquí debemos hacer hincapié en que es el propio sistema el que "descubre" los patrones subyacentes en los datos y a partir de ellos define los diferentes grupos. Los programadores especifican la cantidad de segmentos a obtener, pero no su naturaleza o contenido.

4.2. Construcción de un modelo de aprendizaje automático. ¿Qué categorías son relevantes?

¿Por qué "sexo", "edad" y "raza" sí, y otros criterios no?, se preguntaba Kingma. Sin estirar demasiado la definición de clase de referencia propuesta por Boorse ("una clase natural

de organismos de diseño funcional uniforme") en un sistema de aprendizaje automático sería posible extender las clases de referencia para incluir las múltiples variaciones existentes en la especie humana más allá del sexo, la edad y la raza.

En esta extensión del número de clases de referencia, se podría considerar cualquier combinación de atributos del cuerpo humano que se pueda medir o describir de manera certera: desde el color de ojos a la densidad ósea pasando por el grueso del cabello, la capacidad pulmonar o el peso. Un sistema de inteligencia artificial podría ser entrenado, entonces, con datos de centenares de atributos antropométricos y su distribución estadística: forma y perímetro del cráneo, perímetro abdominal, alineación de las extremidades, hidratación y coloración de la piel y mucosas, etc. Otra posible manera para entrenar este sistema además de los datos antropométricos sería utilizar datos de signos clínicos de personas sanas y enfermas.

Para desarrollar el modelo y poder sacar conclusiones probabilísticas acerca de casos particulares, el sistema necesitaría, además, poder computar los parámetros estadísticamente normales en función de los diferentes atributos. Lo mismo que Boorse hacía con 3 criterios gracias a la inteligencia artificial podría ocurrir con cientos de miles de parámetros, al estilo de la propuesta de Anderson (2008)

Si bien para un entrenador humano procesar cientos, o miles, de atributos combinados resultaría en una complejidad exponencial inasumible, el aprendizaje automático no tendría ningún problema para lidiar con esta cantidad de datos. Al fin y al cabo, una de sus principales características es ser capaz de procesar un enorme volumen de datos.

Tal como hemos explicado arriba, para generar resultados necesitaríamos primero etiquetar los datos de entrada para que el sistema pueda desarrollar un modelo a partir de ellos y logre asignar (de manera probabilística) una etiqueta de salida para un nuevo valor. Resulta evidente que esto no resolvería satisfactoriamente el problema señalado por Kingma. Seguiríamos detectando subjetividad a la hora de definir las clases de entrenamiento. ¿Por qué estos signos y no otros? podríamos seguir preguntándonos. Ahora, en vez de tener tres criterios sin justificación previa, simplemente tendríamos muchos más.

Es posible imaginar una respuesta a esta pregunta. Aunque en este caso el aprendizaje automático no nos ofrecería una justificación en sentido estricto de porqué unos criterios sí y no otros, sí que nos permitiría, al menos en teoría, hacer evaluaciones de salud razonablemente neutras. Dada la enorme cantidad de clases que se podrían definir, sería concebible realizar evaluaciones de salud considerando clases de referencia *elegidas al azar*. Trabajar con cualquier clase de referencia basada en cualquier tipo de datos podría quizás acercarnos a evaluar el estado de salud de un individuo de una manera no normativa.

¿Cómo podríamos llevar esto a cabo? El sistema podría ser entrenado con los datos antropométricos, de la historia médica y de los signos clínicos de personas etiquetadas como sanas o enfermas, pero en este caso utilizaríamos solo una fracción de los datos disponibles, variando de manera aleatoria qué signos clínicos se toman en cuenta y cuáles se dejan fuera. Para intervenir aún menos podríamos dejar que sea el propio sistema, que "decida", antes de comenzar el entrenamiento y de manera aleatoria, qué datos son

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

tenidos en cuenta en la construcción del modelo. Así, el sistema podría incluir los datos acerca de la inflamación o la normalidad de los ganglios linfático, pero dejar fuera los datos acerca de la tensión arterial o el perímetro craneal. Esta técnica se conoce como "muestreo aleatorio" (*random sample*) y se suele utilizar para reducir complejidad computacional, pero no hay nada que impida utilizarla para hacer que la elección de clases esté menos determinada por juicios de valor.

Sin embargo, esto no nos aleja del todo de la influencia normativa. Incluso aquí podríamos seguir preguntando, ¿por qué determinar clases de referencia en base a signos clínicos y no a otro tipo de datos? La elección de utilizar datos clínicos es en sí misma una decisión basada en juicios de valor, independientemente de cuán razonable sea.

Podríamos imaginar, entonces, que el sistema sea entrenado con datos de *cualquier* otro tipo (por ejemplo, las notas del bachillerato, las contribuciones a redes sociales o las declaraciones de impuestos). Lo único que sería necesario desde el punto de vista del aprendizaje automático es que estos datos de entrenamiento estén etiquetados como asociados a una persona sana o enferma y que los volúmenes de datos sean enormes. Una vez entrenado el modelo, el sistema podría asignar una etiqueta de salud o enfermedad a un nuevo valor.

A primera vista, parecería que hemos logrado razonablemente eliminar la subjetividad en la elección de las clases de referencia. Sin embargo, emerge un segundo problema: el de la circularidad. De la misma manera que nuestra *app* necesita aprender qué es un colibrí y qué un tero-tero para poder etiquetar una imagen de un nuevo pájaro, nuestro sistema necesitaría aprender qué es salud y qué es enfermedad. Pero para entrenar al sistema mediante datos etiquetados como "sano" o "enfermo" se precisa de una concepción de salud y enfermedad previa, lo cual viola manifiestamente la propia no-normatividad que se pretende conseguir.

Este breve análisis nos indica que el aprendizaje automático supervisado puede servir para hacer evaluaciones de salud y enfermedad en base a grandes volúmenes de datos una vez definidas las clases de referencia adecuadas, pero no parece ser de ayuda para obtener estas clases de manera no normativa.

V. Concepciones previas, opacidad y sesgos en aprendizaje automático

En esta sección exploraremos si mediante el aprendizaje no supervisado es posible determinar clases de referencia en base a elementos puramente empíricos. Es claramente posible realizar agrupamientos de manera no supervisada. Lo central aquí es si estos agrupamientos pueden realizarse sin intervención de juicios de valor y otros elementos subjetivos que saboteen metodológicamente el naturalismo de Boorse.

Uno de los usos más comunes en aprendizaje no supervisado es el "perfilado" (*profiling*). El perfilado es el desarrollo de modelos estadísticos a partir de *big data* con el objetivo de detectar patrones o estructuras presentes en los datos que no hayan sido previamente hipotetizadas. Mediante el perfilado se pueden hacer predicciones sin necesidad

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

de utilizar modelos causales u otras explicaciones teóricas. Actualmente encontramos este tipo de técnicas con bastante frecuencia. Los sistemas de recomendación de la plataforma Netflix, por ejemplo, incluyen aprendizaje no supervisado para generar y refinar agrupamientos que luego son utilizados en modelos de aprendizaje supervisado (Gomez-Uribe y Hunt, 2016, pp. 2-6). Resulta concebible pensar que el aprendizaje no supervisado y las técnicas de perfilado se podrían utilizar para desarrollar "clases de referencia". Al fin y al cabo una clase de referencia bien puede verse como un tipo de perfil ("mujer blanca de 35 años"). Por ahora no estamos interesados en determinar si estas clases de referencia serían realmente útiles en un sentido operativo, sino si sería posible generarlas sin intervención de elementos normativos.

Podríamos también plantearnos un análisis puro de relevancia estadísticas: meter todos los datos que seamos capaces de recolectar, desatar el *data deluge* y filtrarlo a través de una serie de análisis estadísticos para establecer en las personas sanas que rasgos son los estadísticamente más numerosos. Pero ello tampoco nos llevaría muy lejos. Recordemos las ideas de Millikan acerca de cómo se define la función biológica de un organismo, órgano o proceso. Para entender qué es un corazón sano necesito establecer primero cuál es la función biológica del corazón. Pero esa función no es un simple dato objetivo que se obtiene de analizar la estructura del corazón y ver cuál es su comportamiento normal estadísticamente hablando.Necesito establecer la historia evolutiva del corazón. Saber por qué la selección natural nos ha puesto un corazón es lo que me permite saber que la función última del corazón es bombear sangre y no hacer ruidos rítmicos. Pero esa historia biológica sólo es accesible a través de una serie de teorías y modelos que inevitablemente conducen a elementos valorativos. Si miramos solo el comportamiento estadísticamente normal del espermatozoide acabaríamos concluyendo que la esterilidad en los hombres es la opción más sana. A continuación veremos el sentido de este fenómeno en la salud desde el concepto de índice de masa corporal.

1. La inevitabilidad de una concepción previa de salud

Consideremos como primer caso un sistema que genera una clase de referencia basada en el Índice de Masa Corporal (IMC). El IMC es una medida empírica que *prima facie* está libre de juicios de valor ya que se trata de un mero indicador numérico en función del peso y la altura al cuadrado de una persona. En base al IMC puede determinarse si una persona tiene un peso saludable, sobrepeso, obesidad o un peso insuficiente. Las fuentes médicas consideran un IMC de entre 18,5 y 24,9 kg/m^2 como "normopeso" (SEEDO, n/d). Las personas con un IMC superior a 25 kg/m^2 son consideradas con sobrepeso y las con un IMC superior a 30 kg/m^2 como obesas (OECD, 2019, p. 39).

Sin embargo, este normopeso no refleja la normalidad estadística sino una normalidad teórica (i.e., normativa) por sobre la cual se espera una mayor mortalidad o morbilidad. En España un 53% de los adultos tiene un IMC superior a 25 kg/m^2. La media en países de la OECD es de 58% (OECD, 2019, p. 45); en términos estadísticos, la obesidad y el sobrepeso son más prevalentes que el "normopeso". Es decir, en España

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

hay más personas con sobrepeso que "hombres de más de 75 años de etnia romaní", que podría ser una clase de referencia adecuada según los criterios de Boorse ("sexo", "edad" y "raza").

Imaginemos, entonces, un sistema que optimiza en función del criterio de normalidad estadística tal como propone la teoría bioestadística. ¿Qué razón habría para no aceptar "personas con un IMC superior a 25 kg/m^2" como clase de referencia?

Una primera respuesta que podemos dar es que con esta clase de referencia sucedería algo similar a lo que ocurriría con una clase del tipo "fumadores habituales": una variada gama de estados considerados enfermedades se volverían estadísticamente normales y deberían dejar de considerarse enfermedades. En el caso de un IMC superior a 25 kg/m^2 como referencia se podría afirmar que la insuficiencia cardíaca, la diabetes tipo 2, el cáncer o la artritis no deberían considerarse enfermedades ya que son *estados normales para la clase de referencia*.

Resulta evidente que rechazaríamos esta perspectiva porque la normalidad estadística no concuerda con nuestras intuiciones más primarias acerca de la noción de salud. La insuficiencia cardíaca o el cáncer son ejemplos paradigmáticos de enfermedad. Y más allá del debate de si la obesidad es o no una enfermedad, hay abrumadoras evidencias científicas para considerar la obesidad como un factor clave de riesgo para la salud y por ello como una condición perjudicial para el organismo.

El "normopeso" no refleja ninguna normalidad estadística, sino que se nutre de reflexiones y modelos fisiológicos teóricos, que como todos los modelos son simplificaciones ideales, no descripciones de una realidad empírica. A la vez, podríamos decir, parafraseando a Kingma (2007, p. 131), que no hay ningún hecho empírico que determine qué "neonatos" o "personas de etnia romaní" sí son clases de referencia apropiadas, pero "personas con un IMC superior a 25 kg/m^2" no lo es. Nuestro rechazo a esta clase de referencia sería inherentemente normativista.

Este ejemplo nos indica que resulta implausible esperar que solo un conjunto de datos empíricos sea suficiente para que un sistema de inteligencia artificial pueda determinar una clase de referencia que sea adecuada y naturalista a la vez.

5.2. La opacidad de los sistemas

Un segundo punto de discusión tiene que ver con la opacidad general de los sistemas de aprendizaje automático, que afecta especialmente a los del tipo no supervisado. Las redes neuronales utilizadas para el aprendizaje profundo están hechas de largos vectores de números que hacen difícil, sino imposible, entender qué lleva a un sistema de aprendizaje automático a tomar las decisiones que toma (Marcus y David, 2019, pp. 57-58). Hasta los programadores tienen problemas para entender el comportamiento de los sistemas que ellos mismos han diseñado Es posible saber que un sistema tiene una ratio de éxito de 95% pero aún así resulta imposible en la mayoría de los casos saber por qué falla el otro 5%.

Las redes neuronales generan valores de salida (*outputs*) que están optimizados para el postprocesamiento de datos, pero no proveen explicaciones. Por este motivo se suele

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

decir que los sistemas basados en ellas son del tipo "caja negra" (*black box*). El problema que esto tiene para el aprendizaje automático como apoyo a la teoría de Boorse es evidente. Para considerar una clase de referencia como no normativa es necesario contar con una justificación que resulte plausible. No es imprescindible que la justificación sea en términos de teorías y modelos fisiológicos, pero los sistemas actuales no ofrecen siquiera explicaciones estocásticas mínimas que expliquen el camino de los datos de entrada a los datos de salida. Estamos dispuestos a aceptar sin mayor inconveniente que el recomendador de Amazon nos sugiera que si nos ha gustado el libro del autor X también nos gustará el de la autora Y, pero a la hora de aceptar como válida una clase de referencia el estándar es más alto.

Por esta segunda razón relativa a su opacidad, los sistemas basados en aprendizaje automático no parecen ser candidatos viables para generar clases de referencia que puedan apoyar la teoría bioestadística.

5.3. Los datos no son neutros

Un tercer punto de discusión es de carácter más general y tiene que ver con la naturaleza de los propios datos. La cuestión involucra profundas y complejas cuestiones epistemológicas y por esta razón, nos conformaremos solo con esbozar y plantear unas dudas razonables acerca de la posibilidad de que el aprendizaje automático sirva para generar clases de referencia y fortalecer la teoría bioestadística de Boorse.

El primer aspecto a tener en cuenta es que los datos no son incorporados directamente de manera neutra en los sistemas como si fueran un espejo de la realidad empírica. Es necesario recolectarlos y tratarlos para que sean legibles de manera computacional. Este primer paso ya implica una reducción de la complejidad del mundo a unos campos en bases de datos. Esta reducción no es neutra, sino que está marcada por valores tales como la eficiencia, la efectividad, el coste-beneficio, etc. Pero podemos señalar más problemas para un aprendizaje automático naturalista.

Un problema común son los errores de representación en la selección de los datos (sesgo muestral), que si bien es un problema grave, es hasta cierto punto tratable. Un ejemplo de este tipo de sesgo lo encontramos en sistemas de inteligencia artificial que buscan asistir a dermatólogos en la detección de cáncer de piel. Estos sistemas exhiben gran potencial y alcanzan niveles de predicción comparable o superior al de dermatólogos (e.g. Esteva et al., 2017; Fink et al., 2020). Un problema con estos sistemas es que son mucho más precisos con pieles blancas que con pieles oscuras, lo cual muy probablemente tiene que ver con la manera en que fueron entrenados para reconocer la enfermedad. Las fotografías utilizadas para entrenar a estos sistemas suelen ser las incluidas en el *International Skin Imaging Collaboration*, que es una base de datos abierta y muy rica, pero mayoritariamente compuesta por imágenes de personas blancas (Adamson y Smith, 2018). Una vez identificado el problema, corregir esta desviación es técnicamente viable. Harán falta más imágenes más diversas y posiblemente un nuevo entrenamiento para ajustar el modelo, pero nada de esto es imposible.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

El lado peliagudo de los sesgos en grandes volúmenes de datos es que incluso cuando los datos son estadísticamente representativos y *prima facie* neutros reflejan normas y valores que, a su vez, hacen que las injusticias estructurales existentes en la sociedad sean perpetuadas y amplificadas.

Los datos incluidos en las bases de datos que sirven para entrenar a los sistemas no son neutros. Están necesariamente adecuados a los sistemas y esquemas clasificatorios en los que son incorporados. Bowker y Star (2000) muestran convincentemente como los sistemas de clasificación dan forma a perspectivas sobre el mundo y a las interacciones sociales. Las categorías y los atributos visibilizan unos aspectos, invisibilizando otros, por lo que nunca son un reflejo *naturalista* de la realidad. Sabemos que los modelos en las ciencias sociales pueden cambiar las coordenadas básicas que describen una vez que estas se convierten en políticas (Blakeley, 2020). Un ejemplo es la manera en que se mide "la economía". Indicadores *prima facie* neutros como el Producto Interior Bruto (PIB), la tasa de desempleo o el ïndice Dow Jones Index se proponen como indicadores relevantes, mientras que otros como la humanidad del trabajo, el impacto de las actividades económicas en el medio ambiente o las desigualdades extremas no son tenidos en cuenta. De esta manera, los datos incluidos en el IMC o el PBI, incluso cuando no son sesgados estadísticamente, tampoco son neutros y objetivos, sino que reflejan los valores culturales, políticos, sociales, estéticos e incluso, quizás, religiosos de una sociedad o de los distintos grupos que la componen.

En otros casos los datos no son neutros a causa de un sesgo evidente. La historia de la medicina muestra una falta estructural de interés por la salud de las mujeres. A modo de ilustración, podemos citar el caso en que se ignoran las enfermedades cuando no afectan a los hombres como sucede con la endometriosis (Huntington y Gilmour, 2005) o que los procedimientos y las terapias sean más adecuadas para los hombres que para las mujeres como sucede con las enfermedades coronarias (Beery, 1995).

Los sistemas de inteligencia artificial también se ven frecuentemente afectados por sesgos estructurales. Hay una vastísima literatura en torno a sesgos relacionados con la raza, el género, la edad, el nivel educativo, las capacidades cognitivas y muchos otros vectores de injusticia (e.g. Benjamin, 2019; Eubanks, 2018).

Ilustremos con otro ejemplo, varios autores (e.g. Blakeley, 2020; Harcourt, 2001) han mostrado convincentemente que el patrullaje de "tolerancia cero" (*zero-tolerance policing*), aunque nominalmente ciego a los aspectos raciales, está basado en categorías como "respetuoso con la ley" o "desorden público". Estas categorías, supuestamente descriptivas, están embebidas de significados normativos. Por ejemplo, los llamados "delitos de guante blanco" no suelen ser integrados en estas categorías, de manera que un estafador como Bernie Madoff no entraría a formar parte de los datos de personas no "respetuosas con la ley". Al ser implementadas en algoritmos, resultan en políticas que tienen consecuencias racistas *de facto*, al convertir a barrios pobres y racializados en objetivos de vigilancia policial (Hinton, 2016).

Por este tercer cuestionamiento en torno a los datos, el aprendizaje automático tampoco parece ser un camino prometedor para contribuir de una manera naturalista a una teoría de la salud.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

VI. Conclusiones

En este ensayo hemos intentado explorar si mediante el aprendizaje automático, la inteligencia artificial puede contribuir a determinar las clases de referencia necesarias para la teoría bioestadística de Boorse y, a la vez, a aportar una justificación no normativa de las mismas.

Hemos examinado cinco problemas que muestran importantes obstáculos para que el aprendizaje automático contribuya a fortalecer la visión naturalista de Boorse. Resulta razonable no limitarnos en esta teoría y sugerir que los problemas señalados podrían ser extrapolados a otros contextos en los que se busque "naturalizar" teorías mediante el uso de aprendizaje automático y *big data*. Así, una noción evaluativa de "Justicia" o "Bienestar" cuya definición busque ser formulada de manera estrictamente naturalista encontraría, también, estos mismos problemas, que incluimos aquí abajo de una manera más general:

1. La normatividad en la determinación de clases de referencia es un problema persistente.
2. El etiquetado requerido para los datos de entrenamiento introduce un elemento de circularidad inadmisible desde el punto de vista naturalista.

En la discusión del aprendizaje no supervisado hemos identificado que:

3. Para evaluar si una clase de referencia determinada mediante *perfilado* es adecuada se necesita una concepción teórica previa (en nuestro caso de "salud").
4. La opacidad actual de los sistemas de aprendizaje automático no supervisado no hace viable una justificación satisfactoria para las clases de referencia requeridas por la teoría.
5. Los datos necesarios para un sistema de aprendizaje automático tienen sesgos estructurales por los que se cuelan, inevitablemente, elementos de normatividad.

A los problemas señalados podríamos añadir uno aún más profundo y general: el carácter inherentemente normativo de la tecnología señalado por multitud de autores (para una introducción, véase Radder, 2009). Quedaría por resolver, entonces, si, y hasta qué punto, es posible fortalecer una teoría naturalista mediante instrumentos normativos, pero esta no es la tarea que hemos intentado acometer en este ensayo.

Probablemente sobre aclarar que los problemas señalados no implican que el aprendizaje automático quede invalidado para el análisis evaluativo de la salud, solo resaltan y explicitan la normatividad presente en él. Resulta fácil de imaginar que *una vez definidas las clases de referencia adecuadas*, el aprendizaje automático puede contribuir de manera notable a estos análisis. Sin embargo, en base a nuestro análisis podemos concluir que tanto el aprendizaje supervisado, como el aprendizaje no supervisado arrastran elementos de normatividad y subjetividad que hacen que la aplicación del aprendizaje automático resulte inviable si el objetivo perseguido es una teoría no-normativa y naturalista en sentido estricto.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Para concluir y recapitular lo que hemos estado presentando aquí de forma sucinta, podemos reconstruir nuestro argumento de la siguiente manera:

a) El mero hecho de usar fórmulas matemáticas y datos supuestamente sin interpretar no garantiza la objetividad de un sistema de inteligencia artificial
b) Cualquier modelo que se quiera aplicar a biomedicina, para tener sentido, debe partir de un concepto específico de salud.
c) Si queremos que nuestro modelo biomédico sea objetivo, la concepción de salud debería ser también objetiva.
d) La concepción naturalista de salud de Boorse es en sí misma insuficiente para garantizar una idea de salud totalmente neutra y objetiva.
e) Ninguno de los modelos actualmente en uso para generar algoritmos de aprendizaje automático puede solventar los problemas que presenta una concepción naturalista de salud.

Podemos así responder a la pregunta que encabeza este artículo de forma negativa: El aprendizaje automático es insuficiente para establecer una definición realmente objetiva de salud que permita eliminar aspectos valorativos en el desarrollo de la biomedicina.

Aunque pueda parecer parcial y limitado, este resultado en realidad es muy relevante a la hora de considerar el sentido de incluir el aprendizaje automatizado en la esfera de las decisiones que afectan el bienestar humano. Qué entendemos por salud es la pieza clave sobre la que gira, no sólo nuestra concepción de salud pública, sino también otras dimensiones de la sociedad como la política, la cultura o la educación.

Establecer estos límites en la intervención del aprendizaje automático en estas esferas no implica, ni mucho menos cerrarles el paso. Más bien al contrario, solo estableciendo cuándo tiene sentido su aplicación y cuando no podremos tener una comprensión clara de para qué sirven y de qué forma pueden ayudarnos.

En el camino, sin embargo, tenemos que renunciar a ese sueño de objetividad generada de forma automática y perfecta por máquinas, y dar la bienvenida a un espacio de contribución persona-máquina haciendo complejos equilibrios conceptuales para establecer sistemas de decisiones que sean más precisos y también más justos.

Bibliografía

Adamson, A. S., y Smith, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology, 154*(11), 1247-1248.

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*. Wired Magazine. Recuperado 15/11/21 de https://www.wired.com/2008/06/pb-theory/

Beery, T. A. (1995). Gender bias in the diagnosis and treatment of coronary artery disease. *Heart & Lung, 24*(6), 427-435.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Cambridge: Polity.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Blakeley, J. (2020). *We built reality: How social science infiltrated culture, politics, and power*. Oxford: Oxford University Press.

Boden, M. (2016). *AI: Its nature and future*. Oxford: Oxford University Press.

Boorse, C. (1975). On the distinction between disease and illness. *Philosophy and Public Affairs, 5*(1), 49–68.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542-573.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almeder (Eds.), *What is disease?* (pp. 1–143). Totowa: Humana Press.

Boorse, C. (2014). A second rebuttal on health. *Journal of Medicine and Philosophy, 39*, 683–724.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Bowker, G. C., y Star, S. L. (2000). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.

Carnap, R. (1998). *Der logische aufbau der welt* (Vol. 514). Felix Meiner Verlag.

Casacuberta, D., y Vallverdú, J. (2014, 2014/01/02). E-science and the data deluge. *Philosophical Psychology, 27*(1), 126-140.

Domingos, P. (2021) *We must stop militant liberals from politicizing artificial intelligence.* The Spectator. Recuperado 15/11/21 https://spectator.us/militant-liberals-politicizing-artificial-intelligence/

Ereshefsky, M. (2009). Defining "health" and "disease". *Studies in the History and Philosophy of Biology and Biomedical Sciences, 40*(3), 221–227.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., y Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Fink, C., Blum, A., Buhl, T., Mitteldorf, C., Hofmann-Wellenhof, R., Deinlein, T., Stolz, W., Trennheuser, L., Cussigh, C., Deltgen, D., Winkler, J. K., Toberer, F., Enk, A., Rosenberger, A., y Haenssle, H. A. (2020). Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *Journal of the European Academy of Dermatology and Venereology, 34*(6), 1355-1361.

Feyerabend, P. (1993). Against method. Verso.

Gammelgaard, A. (2000). Evolutionary biology and the concept of disease. *Medicine, Health Care and Philosophy, 3*, 109-116.

Gomez-Uribe, C. A., y Hunt, N. (2016). The Netflix recommender system. *ACM Transactions on Management Information Systems, 6*(4), 1-19.

Harcourt, B. (2001). *Illusion of order: The false promise of broken windows policing*. Cambridge: Harvard University Press.

Heaven, W. D. (2020). *La IA de plegamiento de proteínas de google resuelve un histórico desafío de la biología*. MIT Technology Review. Recuperado 20/01/2021 de https://www.technologyreview.es/s/12935/la-ia-de-plegamiento-de-proteinas-de-google-resuelve-un-historico-desafio-de-la-biologia

Hinton, E. (2016). *From the war on poverty to the war on crime*. Cambridge: Harvard University Press.

Huntington, A., y Gilmour, J. A. (2005). A life shaped by pain: Women and endometriosis. *Journal of Clinical Nursing, 14*(9), 1124-1132.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Kingma, E. (2007). What is it to be healthy? *Analysis, 67*(2), 128–133.

Kingma, E. (2014). Naturalism about health and disease: Adding nuance for progress. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 39*(6), 590-608.

Kosinski, M. (2021, 2021/01/11). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports, 11*(1), 100.

Kovács, J. (1998). The concept of health and disease. *Medicine, Health Care and Philosophy, 1*, 31-39.

Marcus, G., y David, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. New York: Vintage.

Millikan, R. (1984). Language, Truth, and Other Biological Categories. Cambridge, Mass: MIT Press..

Murphy, D. (2015). *Concepts of disease and health*. Recuperado 15/04/2016 de http://plato.stanford.edu/archives/spr2015/entries/health-disease/

Nordenfelt, L. (2007). The concepts of health and illness revisited. *Medicine, Health Care and Philosophy, 10*, 5-10.

Nordenfelt, L. (2016). A defence of a holistic concept of health. In G. É. (Ed.), *Naturalism in the philosophy of health. History, philosophy and theory of the life sciences* (pp. 209-225). Dordrecht: Springer.

OECD. (2019). *The heavy burden of obesity: The economics of prevention*. Paris: OECD Publishing.

Radder, H. (2009). Why technologies are inherently normative. In D. Gabbay, P. Thagard, & J. Woods (Eds.), *Handbook of the philosophy of science* (pp. 887-921). Amsterdam: Elsevier.

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.

SEEDO. (n/d). *IMC*. Recuperado 19/01/2021 de https://www.seedo.es/index.php/imc

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vorvick, L. (2019). *Pulse*. U.S. Department of Health and Human Services. Recuperado 10/02/2021 de https://medlineplus.gov/ency/article/003399.htm

Wang, Y., y Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246-257.

IUS ET SCIENTIA • 2022
Vol. 8 • Nº 1 • pp. 35-56
ISSN 2444-8478 • http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03