UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

EL PROBLEMA EPISTEMOLÓGICO DE LOS BIG DATA EN LA PRODUCCIÓN DE CONOCIMIENTO CIENTÍFICO

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Filosofía para optar al grado y título de Maestría Académica en Filosofía

SERGIO MARTÉN SABORÍO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2023

DEDICATORIA

A Cristina, a mi madre y a mi padre.

AGRADECIMIENTOS

La investigación no hubiera culminado nunca de no ser por el apoyo y la paciencia de amigos, amigas, colegas, tutores y lectores que me han ofrecido lo necesario y más. Agradezco a todos quienes tuvieron discusiones apasionadas e interesadas con mi persona sobre el tema concerniente. Agradezco en especial a Sergio Rojas, Mauricio Molina y Javier Trejos por sus excelentes observaciones y apuntes sin los cuales, el texto que sigue no sería interpretable del todo. También agradezco a Cristina, Valery, Federico, Leonardo, Mario y Carolina por darme, en conversaciones cotidianas y poco formales, perspectivas que antes me eran invisibles. Agradezco también a mis padres, cuya insistencia por completar el camino académico —que aún no acaba— no me ha dejado detenerme por completo, por más lento que sea el viaje.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Filosofía de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en Filosofía."

Dr. Mario Salas Muñoz
Representante de la Decana
Sistema de Estudios de Posgrado

Dr. Sergio Rojas Peralta
Director de Tesis

Dr. Mauricio Molina Delgado
Asesor

Dr. Javier Trejos Zelaya
Asesor

Dr. Mario Solís Umaña
Director del Programa de Posgrado en Filosofía

Sergio Martén Saborío
Candidato

TABLA DE CONTENIDOS

RESUMEN

La ciencia de datos ocupa un puesto de importancia en las ciencias actuales debido a los avances en *Machine Learning* y en la recolección y almacenamiento de grandes cantidades de datos —*Big Data*—, los cuales han posibilitado un nivel más alto de precisión en muchas disciplinas, además de que han hecho cognoscibles ciertos objetos que a simple vista no lo eran. Sin embargo, la razón para la distribución de pesos y sesgos en algunos modelos de *Machine Learning* es incognoscible por su complejidad, lo cual, en el contexto de la producción de conocimiento científico, significa omitir lo explicativo de las teorías. En la investigación se muestran las razones epistemológicas que fundamentan el problema de la opacidad epistémica. Luego, se analiza el área emergente de la *inteligencia artificial explicable* (XAI), un intento de evadir la opacidad epistémica a través de explicaciones locales de los modelos opacos, y se muestran sus alcances y limitaciones. Finalmente, se argumenta que la explicación no es accidental en la ciencia, sino que epistemológicamente debería ser parte esencial de ella. Se muestra que, mientras que el uso de modelos complejos de *Machine Learning* presenta problemas *sui generis* en su uso en las ciencias, estos no atentan directamente contra los métodos científicos actuales. La explicación del funcionamiento de los modelos complejos, mientras que no permite una comprensión absoluta de la razón de cada parámetro, puede abarcar lo suficiente —siempre que se aplique el modelo de explicación adecuado, comprendiendo sus limitaciones— como para permitir la producción legítima de conocimiento científico.

ABSTRACT

Data science occupies an important spot among current sciences due to its development of *Machine Learning* tools and the collection and storage of increasingly big amounts of data —*Big Data*—, all of which have led to a higher level of precision in a lot of different disciplines, as well as they have made certain objects cognizable which at first glance seemed unknowable. However, the reason for why weights and biases are distributed as they are in some complex *Machine Learning models* seems still to be unknowable precisely because of its complexity. This, in the context of scientific knowledge, could imply the need to omit explanations from theories. Here, we show the epistemological reasons that ground the problem of epistemic opacity. We then analyze the emergent area of *explainable AI* (XAI), an attempt to avoid epistemic opacity through local explanations of opaque models, to make clear its scope and limitations. Finally, we argue that explanation is not accidental to science, but that, epistemologically, it should be an essential part of science. The use of complex models of *Machine Learning* for scientific purposes comes with a *sui generis* set of problems, which do not, in fact, go against current scientific methods or practices. The explanation of how these complex models work, while it does not allow for an absolute comprehension of the reason for *each* specific parameter, may comprise enough to allow for legitimate production of scientific knowledge —at least as long as the choice of model is reasonable and rational and the explanation model chosen is adequate and its limitations are understood—.

LISTA DE ABREVIATURAS

Abreviación	Significado
ANN	*Artificial Neural Network* (Red neuronal artificial)
BD	*Big Data*
DL	*Deep Learning* (Aprendizaje profundo)
DNN	*Deep Neural Network* (Red neuronal profunda)
IA	*Inteligencia Artificial*
IML	*Interpretable Machine Learning* (*Machine Learning* interpretable)
ML	*Machine Learning*
XAI	*Explainable Artificial Intelligence* (Inteligencia artificial explicable)

INTRODUCCIÓN

Conforme las técnicas de producción de conocimiento científico se han ido refinando, el conjunto de lo que se considera cognoscible también ha ido en crecimiento. Para Aristóteles la geometría era una de las pocas cosas que podían producir conclusiones con evidencia certera y precisa. La política, lo sociológico, lo ético y otras cuestiones de mayor complejidad debían contentarse con bosquejos y verdades a medias (*Ética a Nicómaco*, I, III, 1094b11-26). Para Galileo, los objetos de estudio que podían disfrutar de esa clase de certeza eran muchos más, ya que ahora se volvía aparente que el lenguaje matemático permitía comprender con certeza el "libro de la naturaleza" en su totalidad (Galileo, 1981, p. 63). Las verdades de la geometría podían aplicarse a los fenómenos físicos para heredar a ellos su evidencia y verdad. Queda la duda de si aquello ético, político y social estaba incluido dentro del libro de la naturaleza del que hablaba Galileo, pero de lo que no hay duda es de que, si se avanza unos siglos más hasta la actualidad, llegar a certezas y verdades precisas en esas materias es una realidad cada vez más plausible. Al menos a esa dirección parecen apuntar las herramientas que brinda el campo de la ciencia de datos, de origen relativamente reciente.

La ciencia de datos ha adquirido paulatinamente un alto nivel de importancia para la sociedad actual. No poco de ello se debe al constante y recíproco avance de dos factores: las técnicas y tecnologías para la recolección y el almacenamiento de datos (los datos como recurso), y las técnicas y tecnologías para el procesamiento y análisis de esos datos. Podemos nombrar los límites extremos de esos dos factores *Big Data* y *Machine Learning* respectivamente[1]. Mediante ellos, aquellas cosas que parecían fuera del alcance de la cognición humana se han tornado asequibles: el uso de simulaciones computacionales ha permitido comprobar y desaprobar teorías en física, lo cual no hubiera sido posible mediante observaciones directamente naturales (Humphreys, 2009); el análisis de información genética es posible gracias a las enormes cantidades de datos que son condición necesaria y que ahora sí pueden ser analizadas (Stephens *et al*., 2015);

[1] Evidentemente también es posible el análisis de conjuntos relativamente pequeños de datos mediante técnicas estadísticas que no se incluyen necesariamente dentro del *machine learning*, pero aquí resaltamos la versión más extrema de esos factores: conjuntos enormes de datos y técnicas complejas y automáticas de análisis.

la sociología y disciplinas aledañas han encontrado en estas herramientas una forma de aumentar la precisión sobre sus conclusiones y teorías (Molina & Garip, 2019).

Sin embargo, el abuso de esas mismas tecnologías en ámbitos empresariales y gubernamentales ha hecho necesario preguntarse si también es posible algo similar en la ciencia. ¿Conllevan los resultados del análisis de grandes conjuntos de datos una mayor oscuridad en sí mismos que aquello que intentan aclarar? El problema del paulatino incremento en el uso de *Big Data* y *Machine Learning* es que en parte se ha dado acríticamente, sin considerar los efectos que una adopción precipitada pueda tener. Entre las consecuencias que se han mostrado, se puede incluir la crisis de la replicación en la ciencia producida mediante *Machine Learning* (Kapoor & Narayanan, 2022) y las crecientes preocupaciones surgidas del problema que será central en esta investigación: la opacidad epistémica —también llamado el problema de la caja negra—.

Los modelos de *Machine Learning* varían en su nivel de complejidad: desde modelos simples, como una regresión lineal, hasta modelos muy complejos, como una red neuronal multicapa, la cual también se engloba en el concepto de *Deep Learning.* Estos últimos están constituidos de tal forma, que comprender *por qué* funcionan como lo hacen para llegar a las predicciones específicas a las que llegan se podría juzgar inescrutable. El asunto es que detrás de esa inescrutabilidad se esconden problemas de sesgo, tanto éticos como técnicos y, en general, una desconfianza en los resultados —aunque precisos— del modelo.

¿Es posible eliminar la opacidad al menos en parte? El campo de *Machine Learning Interpretable* es un área bastante reciente de la ciencia de datos que busca lograr este fin. Por ello, es importante considerarla como parte de la respuesta a la pregunta central de esta investigación: ¿ha producido la adopción de las herramientas de *Big Data* y *Machine Learning* en la ciencia un cambio fundamental en la práctica científica misma? A la par de esta pregunta, se encuentran otras que ya se han hecho, pero que han encontrado respuestas muy variadas a lo ancho de la literatura especializada: ¿significa la ciencia centrada en datos un fin para la teoría? ¿Debemos hacer un intercambio entre la interpretabilidad y la capacidad de de la ciencia, y su precisión[2]?

Para dar una respuesta sistemática a estas cuestionantes y otras que surgirán a lo largo de la exploración sobre el tema, esta investigación constará de tres capítulos. En el primero, *Análisis de*

[2] Sobre el concepto de precisión se darán algunas aclaraciones en el primer capítulo debido a las diferentes acepciones con las que se puede comprender.

datos, Big Data y la complejidad de los modelos, se dará una exposición de los conceptos principales relacionados con el campo de la ciencia de datos, así como se planteará un par de ilustraciones para aportar claridad en los siguientes capítulos en que se desarrolle la propuesta propia. Estará constituido por cuatro subsecciones que buscan respectivamente: (1.1) Definir lo que se enmarca dentro de la ciencia de datos, (1.2) definir el concepto de *Machine Learning*, (1.3) ejemplificar los modelos de *Machine Learning* revisando varios de diversos niveles de complejidad y (1.4) definir con precisión el concepto de *Big Data*.

El segundo capítulo, *Cajas negras e interpretabilidad de modelos: una aproximación epistemológica*, planteará con detalle el problema de la opacidad epistémica, al tiempo que explorará una de las principales potenciales soluciones que se están desarrollando: los modelos de *Machine Learning Interpretable*. Este último será problematizado y se mostrarán sus límites. Para ello, cuenta con otras cuatro subsecciones: (2.1) definirá la opacidad epistémica, (2.2) planteará las características principales del *Machine Learning Interpretable*, (2.3) ilustrará lo dicho con ejemplos de dos modelos de interpretación específicos y (2.4) determinará los problemas fundamentales de esta área de la ciencia de datos.

Por último, el tercer capítulo intitulado *¿El fin de la teoría? La investigación basada en datos y los límites de la ciencia* llega a las respuestas de las preguntas planteadas más arriba sobre las bases teóricas adquiridas en los capítulos anteriores. Cuenta con tres subsecciones: (3.1) define los conceptos epistemológicos centrales para poder responder adecuadamente a las preguntas, (3.2) presenta el rol de *Machine Learning* y *Big Data* en la práctica actual utilizando varios ejemplos y (3.3) concluye que, a pesar de haber un cambio en la forma de hacer ciencia, este cambio no es fundamental ya que no afecta las bases de lo que se considera necesario para la ciencia. La teorización sigue intacta, así como la explicatividad de la ciencia. A lo largo del capítulo, se problematiza el uso indiscriminado, acrítico y apresurado de estas herramientas, pero no por ello se niega su utilidad y el progreso que significa para las diferentes disciplinas que lo adoptan.

Una nota antes del inicio del desarrollo: algo recurrente a lo largo de esta investigación, fue la problemática que surge de una tarea interdisciplinaria como la que aquí se lleva a cabo. Se trata de una combinación entre filosofía (epistemología), estadística y ciencias de datos y computacionales. Cada una de estas disciplinas emplea términos generales como “modelo”, “interpretación”, “datos”, “información”, “precisión”, “incertidumbre” y “sesgo” de maneras muy diferentes. Lo que se ha intentado a continuación, ha sido explicitar las definiciones sobre las que

se fundamenta la investigación para por lo menos partir del mismo material de razonamiento. Se entiende que no todas las personas de todas las disciplinas mencionadas estarán satisfechos con el uso dado de esos y otros términos, pero en ese caso sería productivo el basarse en cómo se definió de manera explícita tal término en este texto y partir de ahí. Parece ser una de las mejores formas de lidiar con uno de los problemas fundamentales de la interdisciplinariedad que parece ser la inconmensurabilidad lingüística. Esto no afectó solo la escritura de la investigación, sino también en el proceso de revisión y lectura de fuentes: se consultaron autores especialistas en una gran variedad de campos relacionados con las disciplinas mencionadas, y cada uno parece utilizar las clases de términos ejemplificadas de maneras muy distintas. Por este motivo, establecer una matriz conjunta *ad hoc* parece ser la única solución, ya que esperar un cambio en todas las disciplinas en la manera en que hacen uso del lenguaje resulta, cuando menos, ingenuo.

Capítulo 1: Análisis de datos, Big Data y la complejidad de los modelos

Para responder a la pregunta que incumbe a esta investigación —si las herramientas complejas de análisis de datos y los *Big Data* cambian la estructura de la producción de conocimiento cualitativamente, y no solo cuantitativamente—, primero se debe sentar el marco con el que se aproximará al tema. Este está compuesto por dos ejes centrales de igual importancia. El primero se construye respondiendo a las siguientes cuestiones: ¿Qué es el *machine learning* (ML) y qué lo diferencia de otras herramientas utilizadas para la producción de conocimiento? Y, ¿a qué nos referimos, precisamente, cuando hablamos de *Big Data* (BD) y cómo interactúa con ML? El segundo eje consiste en la perspectiva epistemológica desde la que se tratará el problema. Este ofrece el sistema que modela cómo producimos conocimiento, mientras que aquel funge como un elemento *prima facie* externo que podría alterar tal *modelo* en alguna medida[3]. Este capítulo y parte del siguiente se dedicarán a construir el primer eje.

Para tales fines, en esta sección se expondrán los fundamentos de la construcción en forma de una presentación y análisis de los conceptos centrales, indispensables para comprender y dar solución a los problemas que con ellos se relacionan. Una primera sección (1.1) se dedicará a esclarecer en qué consiste la disciplina general de la ciencia de datos. Luego, en (1.2) se ofrecerá una explicación de la particularidad de las herramientas conocidas como inteligencia artificial (IA), específicamente las relacionadas con ML y, (1.3) aún más específicamente, las regresiones lineales, los árboles de decisión y las que hacen uso de redes neuronales artificiales (ANN, por sus siglas en inglés), con el fin de mostrar tres niveles distintos de complejidad en modelos de ML. Posteriormente, en (1.4) se sentará una definición adecuada del concepto de Big Data, ya que se utiliza con muchas acepciones distintas en la literatura sobre el tema en general, especialmente por ser una expresión en boga e imprecisa. Con este propósito, se explorarán diversas respuestas técnicas ofrecidas recientemente a la pregunta por la esencia de los Big Data (Crawford *et al*., 2014; De Mauro *et al*., 2015 y 2016; Kitchin, 2014; Leonelli, 2014; entre otros). La última sección (1.6) estará dedicada a exponer propiamente el problema de la opacidad epistémica, conocido también

[3] Aquí se utilizan los conceptos de *modelo* y *sistema* haciendo referencia a la exposición del método de los niveles de abstracción que realiza Floridi (2008 y 2011, cap. 3), explicado de forma sintética en la próxima sección. Cuando se utilice en este sentido, se marcará en itálicas. Más adelante se hace uso de otros sentidos del concepto de "modelo", los cuales se aclararán cuando sea oportuno y no serán marcados con itálicas.

como el problema de la caja negra (Brendel *et al*., 2018; Castelvecchi, 2016; Koh & Liang, 2017; Krause *et al*., 2016; Lipton, 2018; Papernot *et al*., 2017; Rudin, 2019; Watson *et al*., 2019), a modo de conclusión del primer capítulo, en la que se combinan todos los conceptos aclarados hasta ese momento.

1.1 Ciencia de datos: tres elementos constituyentes

La evolución de la ciencia de datos en años recientes apunta a que la cantidad y la calidad no son dos factores completamente opuestos entre sí. A pesar de que una perspectiva común en la cultura popular es que estos términos se conciben como elementos de un intercambio inversamente proporcional, donde el aumento de la cantidad disminuye o dificulta en igual medida la calidad, en ocasiones no es este el caso. En efecto, lo meramente cuantitativo muchas veces puede implicar una transformación al nivel cualitativo, y no necesariamente con proporcionalidad inversa. Prueba de ello es la *cantidad* masiva de datos que se pueden recolectar, almacenar y procesar en la actualidad, lo cual ha dado paso a una vertiginosa transformación *cualitativa* sobre su utilidad y sobre la forma en que se manejan. Este es el caso, como se dijo al principio, de la ciencia de datos[4].

La ciencia de datos es una disciplina que empezó a formarse desde hace más de cinco décadas. En 1962, John Tukey proponía una mezcla entre la estadística matemática, de corte más teórico y principalmente basada en la teoría de la probabilidad, y las técnicas para inferir lo general de lo particular *en la práctica*, conocidas como inferencia estadística:

> Después todo, he llegado a sentir que mi interés central está en el *análisis de datos*, el cual asumo que incluye, entre otras cosas: procedimientos para analizar datos, técnicas para interpretar los resultados de esos procedimientos, formas de planear la recolección de datos para hacer su análisis más fácil, más preciso o más exacto, y toda la maquinaria y resultados de la estadística (matemática) que aplican al análisis de datos. (Tukey, 1962, p. 2)[5]

La propuesta original del autor es que el análisis de datos debía considerarse más como una *ciencia* que como un *saber* meramente *teórico*. La ciencia y el saber teórico se distinguen apelando a la

[4] Sobre el tema del intercambio entre cantidad y calidad y cómo este afecta los métodos de producción de conocimiento científico se hablará en el capítulo 3, sección 3.

[5] Todas las traducciones de inglés a español son propias, a menos que se indique lo contrario.

practicidad de los resultados: la matemática, por ejemplo, no es una ciencia desde esta perspectiva, sino un saber teórico, ya que "su estándar de validez ulterior es un tipo de consistencia lógica y demostrabilidad sobre la que se ha acordado" (p. 6). La ciencia, más bien, tiene como estándar de validez ulterior la "prueba de la experiencia" (p. 5). De esta forma, el análisis de datos se debe centrar en el alcance y la utilidad, en vez de la seguridad (demostrabilidad matemática), lo cual conduce a una anuencia moderada a errores y al uso de la matemática como base para *juzgar*, y no como base para *probar* o *validar* (p. 6). Más aún, el autor reconoce las computadoras como una parte vital en muchos ámbitos de esta ciencia, puesto que hacen posible el análisis de conjuntos grandes de datos que, sin estas máquinas y su capacidad operacional, sería imposible de otra manera (p. 64). Tal reconocimiento se da incluso antes de que se generalizara el uso de la informática en la academia. Pero lo anterior no señala una partición absoluta entre la estadística clásica y el análisis de datos, sino que más bien plantea una evolución de la estadística misma. Es decir, el análisis de datos sigue siendo estadística, pero, a diferencia de la estadística teórica, se preocupa más por las inferencias *prácticas* que puede realizar, para lo cual hace uso también de la *inferencia estadística*, una de las partes centrales de aquella. En síntesis, se trata de una ciencia cuyo factor distintivo es el estar intrínsecamente ligada al uso y la creación de técnicas y tecnologías computacionales que la posibilitan. Como se verá en el tercer capítulo, es posible (y deseable) producir teoría a base de esas inferencias prácticas, que es lo que la diferencia de ser idéntica a las herramientas que utiliza.

Por varias décadas, la aplicación de la estadística que planteó Tukey se siguió practicando y evolucionó en varias ramas similares, pero con sus particularidades y nombres distintos, que hacen uso de distintas técnicas de la inferencia estadística, la estadística matemática y las ciencias computacionales: minería de datos, descubrimiento de conocimiento, descubrimiento de conocimiento en bases de datos, entre otros. Pero, como Tukey, hubo varios pensadores muy influyentes que tenían ideas similares sobre esta nueva perspectiva científica. En Francia, Jean-Paul Benzécri era defensor del análisis de datos (*l'analyse des données*) como una disciplina que, mientras que dependía de la matemática en algún grado, no se reducía solo a ella y sus técnicas, sino que se apegaba más a la práctica de las distintas disciplinas que la fueran a utilizar:

> Esta perspectiva es filosófica: no se trata de traducir directamente en términos matemáticos el sistema de conceptos de una disciplina particular para enlazarlos en las ecuaciones de un modelo, ni de aceptar los datos tal y como son producidos; sino de desarrollarlos en una síntesis profunda, que descubre nuevas entidades y las relaciones simples entre ellas.

> En el cálculo diferencial, las situaciones experimentales son divididas admirablemente en componentes simples, los cuales han sido traducidos en una igual cantidad de leyes fundamentales. Nosotros creemos que le está reservado al análisis de datos el expresar adecuadamente las leyes de aquello que, complejo por esencia (ser vivo, cuerpo social, ecosistema), no puede ser disecado sin perder su naturaleza. (Benzécri, 1983, p. 11)

Como evidencia Benzécri, el fin de esta disciplina es, principalmente, encontrar leyes en sistemas esencialmente complejos. Diluir esos sistemas complejos en otros más simples, significaría acabar con los sistemas mismos, por lo que solo una disciplina que pueda abarcarlos en su complejidad sería adecuada. Esta disciplina es, según Benzécri, el análisis de datos.

Finalmente, a inicios del siglo XXI se consolidó con el nombre que ahora se utiliza mayormente: ciencia de datos. Sin embargo, no se puede decir que esos nombres anteriores no señalaban directamente a la esencia de la disciplina. Lo que tienen en común es que muestran que el centro de ella se ubica en el buscar y encontrar valor en determinados conjuntos de datos. “Minería”, “descubrimiento” y “análisis” son términos que apuntan a esto último. A fin de cuentas, esa es la definición más simple que se puede dar de lo que la constituye: se trata de una ciencia que tiene como finalidad el trocar datos puros —crudos— en algo de valor (Van der Aalst, 2016, p. 10), o, en otras palabras, el estudio de la extracción *generalizable* de conocimiento a partir de conjuntos de datos (Dhar, 2013, p.64), para lo cual, como ya percibía Tukey (1962), son vitales las tecnologías, los marcos y los algoritmos que es posible aplicar solo gracias a las computadoras (Bahga & Madisetti, 2016, p. 22).

Podría pensarse que “ciencia de datos” es solo un sinónimo de “estadística”. Después de todo, la estadística utiliza conjuntos de datos para extraer conclusiones generales, no exactas. Es decir, también transforma los datos puros en algo de valor. Pero, a pesar de lo que aparentan, se trata de dos disciplinas no trivialmente distintas. La estadística y la ciencia de datos comparten secciones de sus conjuntos, pero ninguna abarca completamente a la otra. La característica fundamental que las distingue es que la ciencia de datos lidia directamente con la *clase* de datos que posteriormente analiza, puesto que los datos de los que se dispone son sumamente heterogéneos y varían entre estructurados y no-estructurados, estos últimos siendo los más, además de que van en incremento constante (Dhar, 2013, p. 64). En otras palabras, la ciencia de datos tiene como una de sus preocupaciones principales el teorizar sobre cómo adaptar clases muy

dispares de datos a modelos más homogéneos y, aún más, teorizar sobre cómo dar estructura a los datos que, en su forma 'cruda' no la tienen del todo, para posteriormente llevarlo a la práctica.

Para comprender lo anterior, es imperioso distinguir entre datos *estructurados* y *no estructurados*. Los primeros vienen con un modelo de datos ya determinado. Esto es, están ya categorizados de forma tal, que una computadora puede utilizarlos directamente (Kiefer, 2016, p. 62). Por ejemplo, información simbólica sobre individuos humanos: edad, número de tarjeta de crédito, cantidad de transacciones realizadas en un mes, tipo de sangre, entre otras. Estas características (atributos) permiten formar conjuntos de individuos que las compartan o que incluyan igual cantidad de individuos con valores distintos para cada característica (variables), o extrapolar correlaciones relevantes. En otras palabras, cuando se define una unidad estadística, que es el elemento base para la tabulación o compilación, se define también el conjunto de atributos desde el que se va a analizar. Cuanto más general sea la unidad estadística, tanto mayor la cantidad de atributos que se pueden considerar. Si se poseen tabuladas las variables de los atributos determinados de análisis, se puede hablar de datos estructurados. Como ilustración, si la unidad estadística es "hombres con presión alta", algunos atributos relevantes podrían ser: edad, salario, peso, altura, entre otros. Todos son atributos cuyas variables están tabuladas.

Si no se conocen las variables de los atributos determinados, y por lo tanto no se han tabulado o no se pueden tabular, o si ni siquiera se sabe qué variables responderían a los atributos determinados, se puede hablar de datos no estructurados. Se trata de aquella información que no tiene un modelo de datos determinado. No tienen (mucha) categorización predeterminada, lo cual los hace de difícil acceso, organización o estructuración, compilación o tabulación. Son unidades estadísticas para las que no hay atributos relevantes mediante los que se puedan clasificar. De hecho, es esta la característica que funciona para diferenciar a la estadística de la ciencia de datos. La primera solo se da con unidades estadísticas acompañadas de atributos definidos y tabulados, mientras que la segunda sí se puede relacionar con unidades estadísticas cuyo modelo de datos no ha sido definido (más bien busca definir ese modelo de datos). Se trata de videos, imágenes, textos, grabaciones de voz, etc., que, más allá de una fecha o un título que los identifique, no tienen atributos más significativos que permitan hacer uso efectivo de ellos. Sin embargo, es preciso aclarar que no se trata de una característica estrictamente binaria: el nivel de estructuración de los datos responde a grados distintos según el nivel requerido de estudio.

Piénsese, por ejemplo, en Twitter. Al hacer un tuit en esta red social, no se le atribuye inmediatamente una variable como "agresivo", "alegre", "depresivo" o "sarcástico" para distinguirlos según el atributo de "sentimiento del tuit". Tampoco se le pide al usuario que lo describa de alguna de estas formas. De este modo, se sabe la fecha en que se hizo el tuit, se sabe quién lo hizo e incluso desde qué parte del mundo lo hizo. Pero no se sabe qué dice el tuit, qué clase de mensaje es, sobre qué tema habla, las afecciones ligadas a él, entre otros detalles que son justamente lo que los hace relevantes para fines específicos. En este caso, se puede entender como una mezcla entre datos estructurados y no estructurados: el tuit en tanto tuit tiene información estructurada (la mencionada arriba: fecha, usuario que lo produjo, etc.), pero el tuit en tanto *texto* —en tanto *lenguaje*— no está estructurado del todo.[6] De otra manera, utilizando la distinción establecida anteriormente, cuando de la unidad estadística en cuestión no se han obtenido las variables que responden al atributo que se requiere, se conoce como un dato no-estructurado.

¿Cómo lidia, entonces, la ciencia de datos con esta clase de información no estructurada? A través del diseño de modelos de ML que automáticamente y de forma sumamente eficiente (en comparación con los humanos) les dan una estructura. Para continuar con el ejemplo anterior: los tuits en cuanto *texto* pueden adquirir etiquetas —variables— según los atributos mencionados, mediante el uso de modelos de ML que se engloban en el área conocida como "análisis de sentimiento (*sentiment analysis*)", y que toman de otros campos como la lingüística y el procesamiento de lenguaje natural (Liu, 2015, p. xi). Aquí se ofrece un pequeño vistazo a lo que se analizará posteriormente: ML como parte "vital" de la ciencia de datos. En efecto, para volver al punto inicial, a diferencia de la estadística, el estudio y la puesta en práctica de cómo dar valor a los datos, incluso sin la estructura relevante, es esencial y distintiva para esta disciplina relativamente nueva. Más adelante se volverá a tocar los conjuntos estructurados y no estructurados de datos, cuando se especifique la importancia de los *Big Data* y su relación con el problema epistemológico que producen.

---

[6] Incluso, se podría analizar esta "mezcla" como, más bien, una diferencia en niveles de abstracción (Floridi, 2011) con respecto a qué tan estructurados están los datos: un nivel de abstracción específico es el de lo que se conoce como los *metadatos* de, por ejemplo, un libro (autor(a), fecha de publicación, editorial, género, etc.), y un nivel de abstracción distinto es el de lo que significa el *texto* del libro (podríamos decir, un nivel de abstracción *semántico*), que responde a preguntas como: ¿cuál es la finalidad del texto? ¿Cuál es la opinión del autor sobre el tema central? ¿Qué moraleja se puede obtener de las peripecias de los personajes principales?, entre otras preguntas de esta clase. Es decir, la unidad estadística parece ser la misma: libros. Pero los diferentes conjuntos de atributos según los que se clasifica esa unidad conforman niveles de abstracción distintos.

Pero, además de la estadística y las ciencias computacionales, una tercera parte de lo que conforma a la ciencia de datos se obtiene de un conjunto cuyos elementos son sumamente distintos entre sí, y que se vuelven más cuantiosos conforme se generan nuevas técnicas de almacenamiento, manejo y procesamiento de datos. En el conjunto mencionado, se encuentran disciplinas tales como la astronomía, la genómica, el análisis de sentimiento —aludido más arriba—, el desarrollo de negocios, la meteorología, el periodismo, los deportes y muchas otras especialidades que pueden extraer valor de la abundancia de datos que es posible recolectar en la actualidad. Mas, aunque se trate de una variable, no es un elemento trivial. La especificidad del dominio con el que se relacionan las dos terceras partes de la ciencia de datos genera muchísimas diferencias en la forma en que se estructuran y se eligen las características relevantes en el análisis. Sin conocimiento adecuado del dominio al que se van a aplicar las herramientas estadísticas y computacionales, no es posible obtener resultados relevantes, efectivos y eficientes (Trovati *et al*., 2015, p. 24).

Por ejemplo, si se intenta aplicar un modelo de ML para la identificación de la presencia de tuberculosis en radiografías de tórax, se podrían dar muchas variaciones indeseadas e imprecisiones en los resultados por el mero hecho de que no se posea la pericia requerida en el campo médico relevante. Los *inputs* que se den al modelo pueden variar muchísimo, desde el tamaño de la imagen, el fondo, el ángulo, etc., lo cual, si se deja desatendido, torna a la herramienta en algo completamente inútil o, peor aún, genera errores que no se identifican a tiempo y que podría tener consecuencias indeseables sobre los pacientes (Sivaramakrishnan *et al*., 2018). Lo anterior solo se puede evitar y solucionar con el conocimiento y familiaridad en el área para la que se utiliza el modelo de ML. Esto atañe a dos actores: quienes crean el modelo, y quienes le dan uso.

La ciencia y el análisis de datos responden a varios tipos de preguntas, las cuales usualmente son clasificadas dentro de los siguientes cuatro grupos (Bahga & Madisetti, 2016; Van der Aalst, 2016)[7]:

a. Análisis descriptivo
b. Análisis predictivo
c. Análisis diagnóstico

[7] En un sentido más general de investigación, se suele hablar de la exploratoria, la descriptiva y la explicativa. Esta clasificación también funciona para la ciencia de datos en la medida en que también se trata de una tarea investigativa. Sin embargo, con la intención de detallar un poco mejor esta ciencia en particular, se han propuesto estas cuatro clases de pregunta.

d. Análisis prescriptivo

Cada uno de ellos tiene métodos, resultados y problemas distintos. El análisis descriptivo responde a preguntas de la clase de "¿qué pasó?" (Van der Aalst, 2016, p. 10), y tiene como propósito principal el tomar datos del pasado y organizarlos y presentarlos de forma compendiada (Bahga & Madisetti, 2019, p. 22). Por ejemplo, se utiliza para computar, con base en datos recolectados anteriormente, cuál fue el promedio de lluvias en un mes específico, o para medir la cantidad de personas de un país específico que ingresaron en una página web en particular. Por lo general, utiliza funciones lineales y estadística básica.

El análisis predictivo va un paso más lejos que el anterior en términos de complejidad. Responde a preguntas de la clase "¿qué ocurrirá (con más o menos probabilidad)?" (Van der Aalst, 2016, p. 10). Es decir, busca predecir eventos —o la posibilidad de que ocurran— mediante la creación de modelos de predicción que se entrenan con datos existentes. Estos son formados a través de las percepciones brindadas por el estudio del problema de alineación (*alignment problem*)[8] y de ciertos procesos de integración[9] y lógica difusa, además de las otras herramientas que utiliza el análisis diagnóstico, como el álgebra lineal, la regresión lineal y el análisis de componentes principales (Bahga & Madisetti, 2019, p. 24). Por ejemplo, se utilizan múltiples descripciones y diagnósticos sobre los datos de lluvias pasadas, para llegar a la posibilidad de que ocurra una tormenta tropical a corto o largo plazo. Si la posibilidad es alta, se puede predecir razonablemente que en efecto ocurrirá. Esta clase de análisis es mucho más compleja que la anterior, ya que requiere de conjuntos mucho más grandes de datos para alcanzar la precisión requerida para producir un nivel de certeza apropiado según el fenómeno que se esté analizando.

El análisis diagnóstico responde a preguntas de la clase "¿por qué pasó?" (Van der Aalst, 2016, p. 10). Utiliza los datos para diagnosticar las razones por las que algo ocurre (u ocurrió), para lo que hace uso de computaciones con teoría de grafos[10] y álgebra lineal. Se puede recurrir a resultados obtenidos mediante análisis descriptivo y simulaciones predictivas para, con base en ellos, obtener un diagnóstico de lo que causa que se obtengan (Bahga & Madisetti, 2019, p. 24). Por ejemplo, se utilizan los promedios de lluvias de varios meses para diagnosticar que está ocurriendo

[8] Más información sobre este problema y su relación con lo presentado en Christian (2020). Tiene que ver con la interrelación y el pareo entre distintos conjuntos de datos.

[9] Más en Geweke (1989). Consiste en, por ejemplo, inferencias bayesianas, cadenas de Markov, etc.

[10] Para más sobre el tema, consultar Bollobás (2013). Debe entenderse también como una serie de algoritmos de compleja implementación.

algún fenómeno natural específico que las causa. Este tipo de análisis tiene un nivel de complejidad mayor que el descriptivo y predictivo. Para la predicción no siempre es necesario conocer la causa. Basta con correlaciones estadísticas en muchos casos. Conocer la causa específica de un evento es un análisis complejo que incluso requiere de experimentación o simulación.

Finalmente, el análisis prescriptivo es el que responde a preguntas de tipo "¿qué se debe hacer para que ocurra?". Requiere de varios modelos predictivos para decidir cuál es la mejor forma de proceder con la finalidad de llegar a los resultados cuya probabilidad se predice. Hace uso de teoría de grafos, problemas de alineación y optimización[11] (Bahga & Madisetti, 2019, pp. 24-25). Por ejemplo, los algoritmos que dan recomendaciones personalizadas de videos en YouTube, que se le presentan al usuario asumiendo que son aquellos videos con la mayor probabilidad de ser vistos basándose en interacciones previas en el sitio.

Como se dijo arriba, estos cuatro tipos de análisis tienen niveles de complejidad variables, los cuales responden al fin con el que se emplean, a la calidad y cantidad de datos disponibles para su aplicación y otros factores relevantes. Independientemente del tipo, no obstante, aún se trate de análisis descriptivo, se requiere de técnicas automatizadas de análisis de datos para llegar al nivel de eficiencia y precisión que los hace factibles en un principio.

Recapitulando, entonces, la ciencia de datos es una disciplina que se encarga de extraer valor de conjuntos de datos, ya sean estructurados o no estructurados, y también de los procesos que hacen posible esa extracción. Combina, por lo tanto, dos elementos: (1) la estadística y matemática, que fungen como las especialidades teóricas que permiten el desarrollo de algoritmos para cumplir con los fines expuestos, y (2) las ciencias computacionales, que *posibilitan* y hacen *factible* en la práctica la aplicación de esos algoritmos, los cuales tienen un nivel de complejidad tal, que un ser humano solo podría —cuando en principio pudiese— llevar a cabo de forma sumamente ineficiente y, en consecuencia, prácticamente inútil. Por eso, se podría entender como una versión de la estadística aplicada. Mas, al ser una disciplina que tiende a la práctica, debe aplicarse a *algo*. Este algo no se encuentra dentro de la misma matemática o estadística —elementos que brindan lo tendiente a la teoría—, ni (necesariamente) dentro de las ciencias computacionales. Por ello, hace falta incluir un tercer elemento a la definición de la ciencia de datos: (3) el dominio específico en el que se aplican 1 y 2. Además, y para finalizar, esta disciplina tiene principalmente cuatro fines: el

[11] Sobre optimización se encuentra información en Calafiore (2014). Se trata de técnicas para la optimización de la precisión de los *outputs* obtenidos con los modelos entrenados.

análisis descriptivo, diagnóstico, predictivo y prescriptivo de los conjuntos de datos disponibles. La siguiente sección especifica el funcionamiento de la clase de herramientas que permiten ejercer tales formas de análisis.

1.2 *Machine learning*: algoritmos, etiquetas y sobreajuste

Para comprender el problema de la opacidad epistémica producida por las herramientas desarrolladas y aplicadas por la ciencia de datos, es esencial comprender su naturaleza en primer lugar. El análisis de conjuntos de datos para encontrar correlaciones y tendencias no requiere necesariamente de la intervención de computadores. Si el conjunto no es muy grande, los datos no muy complejos y no muy interrelacionados, y si el algoritmo a ejecutar es lo suficientemente simple, entonces un individuo o un grupo de individuos pueden llevar a cabo la tarea. Después de todo, un algoritmo no es más que una serie de instrucciones sobre lo que se debe hacer con cierto *input* para obtener cierto *output* (Alpaydin, 2010, p. 1). Tal vez se utilicen computadores, pero no para llevar a cabo estos algoritmos, sino para almacenar y organizar el conjunto de datos. O tal vez, incluso, se utilicen para los algoritmos, aunque estos pudieran en principio ser realizados por una persona. Pero, como se vio más arriba, lo cualitativo no lleva necesariamente a perjudicar la calidad, y puede más bien fortalecerla.

Cuando esa serie de instrucciones es suficientemente compleja, se vuelve muy impráctico ejecutarla "manualmente". Incluso cuando los conjuntos de datos son relativamente pequeños, y el algoritmo en el que se introducen relativamente simple, la eficiencia, velocidad y precisión con la que los procesa una computadora supera con creces la de un individuo humano. Ahora bien, hay cierto punto en el que los datos son de tal naturaleza y de tal cantidad, que ejecutar un algoritmo que obtenga la clase de *output* deseado se vuelve una tarea de extrema complejidad o, incluso, imposible para un ser humano sin las herramientas relevantes.

Es decir, hay dos funciones generales que pueden cumplir los computadores en relación con los algoritmos: (i) su ejecución y (ii) su producción. Como resultado, existen cuatro distintas combinaciones: (1) algoritmos ejecutados y producidos por humanos (sin herramientas, o con herramientas cumpliendo roles secundarios). Por ejemplo, una simple función lineal puede ser tanto producida como ejecutada por un humano. (2) Algoritmos ejecutados y producidos por computadores, pensando, por ejemplo, en el área llamada "síntesis de programas", donde se

utilizan los mismos computadores para crear algoritmos que a su vez otros programas ejecutan. Es un área reciente y no es de uso común, pero existe. (3) Los algoritmos ejecutados por computadores y producidos por humanos, como una red neuronal multicapa, cuyo algoritmo no es tan complejo en producción, pero sí en ejecución. Y (4) algoritmos ejecutados por humanos y producidos por computadores, de lo que no hay ejemplos útiles. De ellas, las combinaciones 2 y 4 son las menos plausibles, dado que, si un algoritmo es lo suficientemente complejo como para tener que ser producido por un computador, un humano probablemente no podría ejecutarlo, y el área de síntesis de programas, que se dedica a la automatización de la producción de algoritmos todavía no llega a los niveles de complejidad requeridos. La combinación 3 es la más común en la práctica de la ciencia de datos. Incluso cuando se relaciona con el análisis descriptivo —que en primera instancia podría parecer el más simple, pero que también puede requerir de cálculos complejos para obtener medidas comunes (media, varianza, simetría, etc.) — son, para fines de la presente investigación, de especial importancia.

La producción automática de estructuras que permitan el uso efectivo de algoritmos es la tarea que se delega al subconjunto de inteligencia artificial conocido como *Machine learning*. En otras palabras, ML consiste en programas computacionales que ejecutan algoritmos para, a través de ellos, producir "automáticamente" un modelo (según las definiciones que aquí se darán a continuación de modelo y algoritmo) determinado. En general, la inteligencia artificial se puede entender como "la habilidad de un sistema para interpretar datos externos correctamente, para aprender de tales datos y para utilizar ese aprendizaje para lograr fines y tareas específicos mediante adaptación flexible" (Kaplan & Haenlein, 2018, pp. 2-3). Esta definición cubre casos como la robótica —el diseño de robots que sean capaces de realizar tareas "inteligentes"— y computadoras como *Deep Blue*[12], que no utilizan necesariamente ML (o, al menos, una versión rudimentaria de ML) para llevar a cabo sus propósitos. Tales serían ejemplos de elementos del conjunto de IA que no estarían, a su vez, dentro del subconjunto de ML. ML es, entonces, una

[12] Computadora desarrollada por IBM entre los 80s y 90s para jugar efectivamente ajedrez. Su modelo algorítmico responde a la rama de IA conocida como GOFAI (*Good Old-Fashioned Artificial Intelligence* — Inteligencia artificial a la antigua, en español—), la cual se cara cteriza por utilizar algoritmos relativamente poco complejos, pero sumamente exhaustivos. *Deep Blue*, por ejemplo, utilizaba una técnica conocida como "búsqueda por fuerza bruta" (*Brute-force search*) (Bernstein, 2005). Se contrapone, por ejemplo, a *AlphaZero*: un programa diseñado para aprender a jugar ajedrez, go y shogi. Utiliza, a diferencia de *Deep Blue*, un modelo de red neuronal al que no se le alimenta con datos de juegos previos, ni con información de libros de aperturas comunes. El modelo se entrena jugando contra sí mismo (al inicio aleatoriamente) con los únicos parámetros siendo las reglas de los respectivos juegos (Silver et al., 2018)

sección de IA que se especializa en el diseño y ejecución de algoritmos que tienen la finalidad de detectar patrones en conjuntos de datos (Murphy, 2012, p. 1), los cuales son utilizados posteriormente para describir, diagnosticar, predecir o prescribir respecto de conjuntos de datos nuevos. Los primeros son los datos de "entrenamiento", que cumplen el rol de la "experiencia" mediante la que el modelo "aprende"[13] para mejorar su desempeño (Mohri *et al*., 2018, p. 1). Se trata de conjuntos de datos que pueden ser estructurados o no estructurados, dependiendo del modelo de ML que se entrena, y que varían en tamaño y calidad.

Aquí cabe una distinción conceptual importante para lo que sigue, entre "algoritmo" y "modelo". Cuando se habla de algoritmos en ML, se habla, como se dijo anteriormente, de un conjunto de instrucciones por las que se pasa un *input* para obtener cierto *output*. Este concepto es, por naturaleza, más general que el de "modelo", ya que hace referencia únicamente a la clase de instrucciones que se giran. En otras palabras, es una serie de instrucciones con los pesos y sesgos todavía no modificados, en su estado por defecto. Un algoritmo es, por lo tanto, un proceso que no depende de haber sido entrenado, o de haber recibido datos. Es la guía que *permite* que haya entrenamiento del todo, para después convertirse en un modelo. El modelo, en este sentido, es el producto que resulta una vez que el algoritmo ha sido entrenado (proceso que se explicará más adelante), o sea, cuando se han establecido los pesos y sesgos *específicos* que se espera produzcan las predicciones buscadas. En adelante, se hará uso de esta distinción de conceptos.

El nivel de estructuración de los datos de entrenamiento influye sobre el modelo de ML que se utiliza para cierta tarea. Modelos más simples requieren de más estructura, que se traduce en *inputs* ya etiquetados con anterioridad (es decir, que se les da una clasificación explícita previa). Esta forma de "aprendizaje" es conocida como "aprendizaje supervisado" (Mohri *et al*., 2018, p.6). Por ejemplo, en visión computacional, si se busca que un modelo pueda determinar si se trata, o no, de un canguro cuando le alimento como *input* una fotografía, debo entrenarlo con un conjunto de fotografías que estén etiquetadas previamente como "canguro" o "no-canguro". Además de estas etiquetas, puede ser importante en modelos relativamente más simples señalar, también explícitamente, qué características de las imágenes son las que permiten distinguir entre "canguro" o "no-canguro". Esto se conoce como selección de variables (*feature selection*), y consiste en

[13] Se colocan estos conceptos entre comillas porque no se están utilizando en el sentido literal. No es lo mismo decir que un humano aprende, que decir que un modelo de ML aprende. Son dos procesos sumamente distintos, pero este lenguaje analógico es el que se utiliza tanto en la literatura especializada, como fuera de ella.

establecer los detalles de los *inputs* en los que se debe concentrar el modelo durante el entrenamiento. Siguiendo con el ejemplo, se puede marcar en los *inputs* de entrenamiento la forma de la cola y las orejas de los canguros (traducida en vectores numéricos que hacen referencia a los pixeles que constituyen la imagen).[14] Esto brinda varios beneficios que son de gran importancia para el problema de esta investigación, entre los cuales están (i) hacer que los modelos sean más fáciles de interpretar para usuarios o investigadores, (ii) hacer más veloz el proceso de aprendizaje del modelo y (iii) permitir un mejor nivel de generalización del modelo (James *et al*., 2013, pp. 24-26).

De ellos, i y iii son fenómenos sobre los que se profundizará más, pues afectan directamente la opacidad epistémica de los modelos aquí investigada. ii se relaciona con una parte del proceso que no requiere de explicación posterior para otros cognoscentes. Es decir, es independiente de si el modelo es transparente u opaco: un modelo puede ser opaco y rápido u opaco y lento, pero también puede ser transparente y rápido o transparente y lento. En ninguna de las cuatro opciones se ve modificada el nivel de opacidad/transparencia. Podemos entender la velocidad de aprendizaje de un modelo como el tiempo que le toma a este, una vez que inicia su entrenamiento, lograr los *outputs* deseados con el margen de error deseado con un conjunto de *inputs* distinto (o modificado) del utilizado para el entrenamiento. Por ello se trata de una característica de suma importancia para la eficacia y utilidad de los modelos. Mas i y iii sí se relacionan directamente con la opacidad.

Sobre i: cuando un modelo utiliza para su entrenamiento datos no estructurados o con poca estructuración (con etiquetas, pero sin selección de variables, por ejemplo), se está perjudicando la capacidad de un individuo de interpretar lo que el modelo está haciendo para llegar de los *inputs* a los *outputs*. Sin este nivel de estructura, no se puede saber tan fácilmente qué detalles del *input* son relevantes para el modelo y por qué, a diferencia de un modelo que utilice selección de variables, en el que esos detalles son conocidos desde un inicio porque se le "mostraron" explícitamente en su entrenamiento (James *et al*., 2013, p. 25). Esta diferencia en interpretabilidad, producto de si se usa o no selección de variables, resulta bastante intuitiva. No ocurre lo mismo en el caso de la generalización.

La generalización —el beneficio iii— es un concepto fundamental para ML (Mohri *et al*., 2018, p. 7), debido a que la utilidad de los modelos depende en última instancia de si sus resultados se pueden obtener en conjuntos de datos generalizados. Es decir, si el modelo producido por ML

[14] Esta estrategia una del conjunto de la inteligencia artificial simbólica. La estrategia que no predetermina patrones al modelo, sino que es tarea del modelo encontrarlos por su cuenta, es llamada subsimbólica.

solo funciona con la información que se utilizó para entrenarlo, mientras que con información nueva su nivel de efectividad decae significativamente, entonces se habla de que no es un modelo generalizable. Este fenómeno se conoce como sobreajuste (*overfitting*), y es uno de los principales problemas contra los que se debe resguardar quien entrena y desarrolla un modelo. Un algoritmo está sobreajustado cuando, más que detectar patrones y características en los *inputs* que posteriormente pueda utilizar para derivar *outputs*, lo que hace es "recordar" esos *inputs*. El sobreajuste se puede dar por varias razones, entre ellas, una selección de variables deficiente o excesiva, y un conjunto de datos de entrenamiento demasiado específico o demasiado extenso. En ambos casos, el modelo se ajusta tanto y tan exactamente al conjunto de datos de entrenamiento y sus variables específicas, que cualquier *input* nuevo que no se encuentre en ese conjunto basta como diferencia para dar resultados imprecisos y parcializados. Por lo tanto, se puede definir de forma más general como el fenómeno en el que los modelos "siguen de cerca los errores, o el *ruido*[15], excesivamente." (James *et al*., 2013, p. 22).

El sobreajuste se puede controlar eligiendo adecuadamente la cantidad y calidad de los conjuntos de entrenamiento, realizando una adecuada selección de variables —lo cual es posible de forma automatizada mediante técnicas computacionales (Liu & Motoda, 2007)— y, en las situaciones en que sea factible, utilizando modelos no demasiado flexibles —más lineales—. Un modelo es más inflexible y lineal, cuando utiliza métodos que permiten trazar un rango limitado de funciones (James *et al*., 2013, p.24). Por ejemplo, el método de la regresión lineal[16] solo puede tener como resultado funciones estrictamente lineales, lo cual significa que, por lo general, es poco plausible incluir todos los puntos de datos del conjunto en ella sin pérdidas de precisión y efectividad significativas. Se habla de modelos más flexibles cuando utilizan acercamientos que ofrecen un rango mayor de tipos de funciones posibles. Los árboles de decisiones y las distintas técnicas que les otorgan mayor precisión a costas de menor interpretabilidad son ejemplos de algoritmos no lineales —aunque estén compuestos de particiones lineales— y, por lo tanto, un poco más flexibles y precisos en cómo calzan los puntos de datos con la función. Las regresiones polinomiales y de *splines* son ejemplos más flexibles que los árboles de decisión.[17]

[15] "Ruido" se entiende como los datos no relevantes incluidos por accidente (en el sentido aristotélico) en la información utilizada para algún fin específico.

[16] Para más detalle sobre este método, revisar Alpaydin (2010).

[17] Revisar Marsh & Cormier (2001) y Ostertagová (2012) para más sobre estos tipos de regresión no lineal.

Aún más flexibles son las redes neuronales artificiales, las cuales son capaces de modelar cualquier función, sea lineal o no lineal (Hornik et al., 1989), gracias a su estructura, ya que es más flexible que la de un árbol de decisiones. En las redes neuronales artificiales es donde más sobresale el problema de la opacidad epistémica —el problema de la caja negra—, debido justamente a esa estructura y los altos niveles de complejidad que permite, como se analizará con mayor detalle en las siguientes secciones y capítulos. Para poder presentarlo con mayor claridad, no obstante, es necesario introducir el funcionamiento de estos modelos de ML con más especificidad.

1.3 De líneas, árboles y neuronas

Esta investigación no busca exponer de forma exhaustiva los diferentes modelos de ML que se aplican en la actualidad. Sin embargo, sí es importante detallar al menos tres modelos específicos, los cuales representan tres niveles distintos de complejidad, puesto que permitirán mostrar la naturaleza del problema central que aquí concierne: la opacidad epistémica que producen los modelos más complejos de ML y su significado para la forma en que se comprende la práctica científica, que depende cada vez más de ellos. Las tres han sido mencionadas en la sección anterior: (1) la regresión lineal, (2) los árboles de decisión y (3) las redes neuronales artificiales.

La regresión lineal *simple* es la más simple y poco flexible de los tres modelos que se van a presentar. Esto se debe, como se dijo arriba, a que el rango de funciones que se pueden trazar a partir de ella es muy limitado. Ahora bien, no se trata de un modelo exclusivamente de ML, y precede a estas técnicas en la historia de la estadística clásica. Sin embargo, se ha adoptado en estos medios por su utilidad para acercar modelos al resultado deseado de forma simple. Su finalidad es el predecir una respuesta cuantitativa (Y) según una única variable predictiva (X). Para ello, se asume que hay una relación lineal aproximada entre Y y X (James *et al*., 2013, p. 61). Matemáticamente, en su versión más simple, se puede representar como:

$$(I) \qquad Y \approx \beta_0 + \beta_1 X$$

β_0 y β_1 son constantes en principio desconocidas que, respectivamente, dan la intercepta y la pendiente de la recta. Se conocen como los *parámetros* del modelo, y se determinan estimándolos mediante los puntos de datos que fungen como la información de entrenamiento. Estos puntos de datos consisten en una serie de *n* pares de observación: una medición de X y una de Y —(x_1, y_1), (x_2, y_2), …, (x_n, y_n)—. Los parámetros son, entonces, los que determinan las diferentes posibles funciones

que se aproximan a representar la distribución de esos pares de observación: se debe encontrar un β_0 y un β_1 tales, que la línea resultante esté tan cerca como sea posible de todos los puntos de datos (James *et al*., 2013, p. 61). Por lo general, la herramienta con la que se calcula ese "tan cerca como sea posible", es la de los *mínimos cuadrados*. Consiste en obtener la diferencia entre los puntos de la función aproximada, y los puntos de los datos (los utilizados para el entrenamiento), para posteriormente sumar sus cuadrados. El resultado de esta suma se debe minimizar tanto como permita, de modo que la diferencia se aproxime a ser la menor posible[18]. Al llegar a ese mínimo, se considera la función como una aproximación más precisa del conjunto de puntos de datos, lo cual permite realizar, ahora sí, predicciones a partir de ella.

Cabe destacar que, cuando los conjuntos de datos son sumamente grandes (cuando *n* es muy grande), no es factible realizar todo este proceso "manualmente". Por eso se requiere de ML para llevar a cabo las operaciones aquí presentadas y, con base en los datos de entrenamiento, determinar —aprender— los parámetros (β_0 y β_1) requeridos para una predicción precisa. Ahora bien, se debe recordar que el modelo aquí presentado se encuentra en su versión más simplificada. Es posible adjuntar otros procesos que hacen al modelo más complejo, pero, a cambio, mucho más preciso y capaz de modelar funciones más variadas. Sin embargo, la base sigue siendo la misma: los parámetros siempre serán lineales[19]. Al lado de la regresión lineal, se puede hablar también de la *clasificación* lineal. La diferencia entre estas dos, es que la regresión se ejecuta sobre valores numéricos, cuantitativos, mientras que la *clasificación* hace referencia a características cualitativas, las cuales adquieren valores *categoriales* (no numéricos). Por ejemplo, el género, las distintas marcas de productos, el tipo de enfermedades padecidas, la religión, entre otras. Estas, por lo general y en tanto variables dependientes, pueden ser programadas para traducirse en valores numéricos de alguna forma (por ejemplo, en pares binarios como el sexo: 1 podría representar "femenino" y 0 "masculino").

En resumen, los modelos de ML de regresión lineal simple tienen una estructura relativamente simple, como su nombre indica, que, por lo tanto, permite que sean fácilmente

---

[18] Para llegar a esta aproximación al mínimo local, se iguala la derivada de la suma a 0. No se profundiza más sobre el tema, porque no hace falta para el punto al que se desea llegar. En caso de interés, se puede consultar Ruder (2016), así como cualquier otro libro introductorio a los temas de ML y regresión lineal, varios de los cuales están citados en este documento. También podría ser de interés el método de descenso por gradiente

[19] Por razones de claridad, la variable (X) puede ser transformada de forma no-lineal (por ejemplo, si se propone $Y \approx \beta_0 + \beta_1 X^2$), y a pesar de ello el modelo y sus predicciones se siguen considerando lineales, ya que sus *parámetros* continúan siendo lineales (James et al., 2013, p. 92).

comprendidos e interpretados por individuos humanos, a pesar de que su ejecución requiera de computadoras para ser eficientes y factibles en la práctica. Los parámetros son pocos y son lineales, además de que solo uno puede modificar cada variable. La posibilidad de trazar y visibilizar la función resultante del modelo, permite otro nivel de acercamiento en términos de interpretabilidad. Las funciones pueden no ser la forma más intuitiva de comprender el funcionamiento de un modelo —una explicación contrafáctica sería, como se verá más adelante, mucho mejor para estos fines—, pero se acercan. Máxime cuando los parámetros que determinan sus pendientes e interceptas son coeficientes únicos (o pocos) y constantes en la función. Este no es el caso en las redes neuronales artificiales, como se verá más adelante.

Subiendo un nivel más en la escala de complejidad, se encuentran los árboles de decisiones, los cuales se introdujeron anteriormente. Se trata de modelos de ML visualmente más interpretables, ya que se pueden presentar gráficamente y se pueden traducir en una serie de reglas, además de que son más intuitivos y fáciles de seguir como modelos de toma de decisión para seres humanos que la regresión lineal misma, a pesar de que puedan ser más complejos y no lineales (Rokach & Maimon, 2008, p. 81). Los árboles de decisión se pueden utilizar con propósitos, tanto de clasificación, como de regresión: se habla de árboles de clasificación y árboles de regresión respectivamente.

Los árboles de clasificación (y regresión) son modelos de ML que dividen el espacio de instancias (los puntos de datos) recursivamente mediante reglas (Rokach & Maimon, 2008, p. 12). Poseen, por lo general, un nodo "raíz" del que solo salen —no llegan— flechas. Los demás nodos reciben una única flecha, lo cual es importante para comprender la interpretabilidad que ofrecen estos modelos. Los nodos de los que sale una flecha se denominan *internos*, mientras que aquellos a los que solo llegan flechas —no salen—, se llaman *terminales* u *hojas*. Cada nodo interno divide el espacio de instancias en dos partes según una función relativa a los valores de esas instancias. Por ejemplo, si el espacio de instancias responde a un conjunto de personas que padecen del corazón, se podría dividir en el nodo raíz según si son mayores o menores de cuarenta años. Esto llevaría a una recursión de la división en el siguiente nodo interno, tanto del lado de los mayores de cuarenta años, como en el de los menores, esta vez determinada por el sexo, y así continuaría hasta llegar a una *hoja*, o nodo terminal, en el que se determinan las proporciones de si sufrirá o no de un infarto a mediano plazo. Este ejemplo es el de un árbol de clasificación. Lo mismo acontece con árboles de regresión, en los que las reglas se expresan como funciones con respecto a los valores numéricos:

si $X \geq S_0$, donde S_0 es un coeficiente determinado manual o automáticamente, entonces ese subconjunto de X pasa a otro nodo interno, en el que vuelve a ser dividido, o a un nodo terminal, en el que acaba la división y se llega a un resultado. La diferencia yace en que los de clasificación dan una variable de respuesta (simbólica), mientras que los de regresión dan resultados continuos por naturaleza (sin respuesta simbólica).

Los valores de las *hojas*, así como los rangos y atributos que causan la división del espacio de instancias, se determinan mediante la proporción de los valores de respuesta en los que los datos de entrenamiento caen dentro del espacio de instancias perteneciente a ese nodo específico. Es decir, y volviendo al ejemplo de más arriba, si se determina que alguien mayor de cuarenta años, de sexo masculino y con hipertensión tiene grandes probabilidades de sufrir un infarto a mediano plazo, es porque las personas de sexo masculino, con hipertensión y mayores de cuarenta años sufren de infartos a mediano plazo en mayor proporción que otros grupos. Es importante notar, además, que los valores por los que se divide el espacio de instancias en cada nodo interno son mutuamente excluyentes. Es decir, cada elemento de ese espacio de instancias completo solo puede estar en uno de los nodos de un nivel determinado, ya sea terminal o interno. Esto es parte importante de la interpretabilidad que facilitan esta clase de modelos.

Ahora bien, si los árboles tienen pocos nodos, son sumamente fáciles de seguir e interpretar. Incluso cuando poseen muchos nodos, debido a que cada instancia particular puede seguir un único camino, el cual es además unidireccional, se mantiene esa interpretabilidad. En otras palabras, es fácil saber lo que está realizando un modelo de árbol de regresión o de clasificación, incluso si las funciones que determinan la división del espacio de instancias son complejas, ya que se puede representar gráficamente y con una forma que se comprende intuitivamente. Hasta sería razonable decir que los árboles de decisión, aunque sean no lineales, son más fáciles de comprender que los modelos de regresión lineal, incluso para alguien sin pericia en el tema. El costo de esta ventaja se observa en su precisión. Los modelos de árboles de decisión, especialmente los simples, tienden a no ser tan efectivos con sus predicciones en comparación con otros modelos de ML. Además, si para lidiar con esto se construye un árbol más profundo, se corre mucho más peligro de sobreajuste, dado que las muchas divisiones del espacio de instancias dejarían resultados excesivamente específicos y, por lo tanto, de poco valor predictivo.

La solución a este problema se encuentra en varios métodos que utilizan como base los modelos de árboles de decisión, pero le otorgan la precisión predictiva que les hace falta —por

ejemplo, los bosques aleatorios (*random forests*), el empaquetado (*bagging*) y la poda (*pruning*)—, entre otras ventajas como la reducción de sobreajuste y mayor versatilidad y escalabilidad. Tales métodos consisten por lo general en la agregación de *varios* árboles de decisión y el promediado de sus resultados (James *et al*., 2013, p. 320). Estas soluciones eliminan en alguna medida la ventaja expuesta. Un modelo de bosques aleatorios es mucho más difícil de seguir que un único árbol de decisiones, ya que se trata de árboles múltiples con estructuras y funciones diversas. Como excepción, cabe aquí señalar el *pruning*, método que más bien elimina secciones del árbol que sean redundantes, para así eliminar el peligro de sobreajuste y, al mismo tiempo, incrementando su interpretabilidad. Por ello se coloca aquí a los modelos de árbol en una posición intermedia de complejidad e interpretabilidad. Por sí mismos son muy fáciles de seguir, incluso más que la regresión lineal, pero cuando se multiplican y se transforman para lograr mayor precisión, su complejidad aumenta significativamente y, con ella, se reduce la interpretabilidad. Sobre el intercambio entre precisión e interpretabilidad que parece existir por los casos de los que se ha hablado, se hablará en el segundo capítulo.

Finalmente, en el nivel más alto de complejidad y, conversamente, en el nivel más bajo de interpretabilidad, se encuentran los modelos de ML que utilizan redes neuronales artificiales (ANN, por sus siglas en inglés). Se encuentran dentro del subconjunto de ML que suele llamarse *Deep learning* (DL), haciendo referencia a las capas profundas de neuronas que conforman gran parte de las redes. Las ANN se construyeron sobre la base de un funcionamiento análogo al de las redes de células nerviosas en los cerebros humanos y de animales no-humanos (Graupe, 2013, p. 1). Tienen como finalidad el resolver problemas “complejos, mal definidos, altamente no lineales, de muchas y diferentes variables y/o estocásticos” (Graupe, 2013, p. 3). Las primeras formulaciones de los principios que deben seguir las ANN datan de 1943, en un artículo escrito por McCulloch y Pitts (1943). Los autores teorizaron que, por existir la ley de “todo o nada” en la activación de las neuronas (haciendo referencia al potencial de acción, donde, si una fibra nerviosa es estimulada, lo hace al máximo y con una única amplitud), es posible traducir los eventos neuronales puramente a un lenguaje de lógica proposicional (McCulloch y Pitts, 1943). Así inicia la representación de redes (biológicas) neuronales mediante redes (digitales) neuronales. De los principios expuestos en ese documento, pocos se utilizan en las ANN de la actualidad (Graupe, 2013, p. 9) debido a que no podían modelar todas las funciones lógicas, específicamente la o exclusiva (*XOR*) (Minsky & Papert, 2017), a pesar de que esa era su finalidad. Desde entonces, se han desarrollado múltiples modelos, tanto de las redes, como de las neuronas que las constituyen.

El modelo de neurona utilizado con más frecuencia en la actualidad es el *perceptrón*, originalmente desarrollado por Rosenblatt (1958), el cual consiste en una aproximación simplificada a la forma de funcionamiento de las neuronas naturales, y que, mediante la inclusión posterior (no de parte de Rosenblatt) de capas ocultas, permite, ahora sí, modelar todas las funciones lógicas. Cabe anotar que el perceptrón original (Rosenblatt, 1958) no era capaz de modelar el XOR. Fue su modificación posterior la que lo permitió. El perceptrón multicapa es una estructura sobre la que se implementan algoritmos de aprendizaje. Consiste en el conjunto de una función que suma todos los productos de los *inputs* que recibe y los pesos (*weights*) —uno de los parámetros del algoritmo—, y otra función que cumple análogamente el rol del potencial de acción que utilizan las neuronas naturales, y se encarga de definir finalmente el *output* de la neurona para determinar si se activa, o no. El algoritmo, antes de pasar por la función de activación, se representa como:

(II) $$z = \sum_{i=1}^{m} w_i x_i + b_0$$

m es el número total de *inputs* que conectan con esa neurona, mientras que x_i es la variable que representa el valor del i-ésimo *input*. Los parámetros del algoritmo son representados por w_i y b_0. w_i representa el i-ésimo valor de los pesos, cuya función consiste en fijar a qué tasa se ve influenciada la neurona ante los *inputs* que recibe: cuanto mayor el valor absoluto del peso, tanto más afectan pequeñas diferencias en los *inputs*, mientras que, si el valor absoluto del peso es menor, no tienen tantas consecuencias. b_0 corresponde al sesgo (*bias*), el cual modifica la intercepta de la función. Una forma de representar el sesgo se da incluyendo en el modelo un *input adicional* cuyo valor x_0 (en vez de iniciar por x_1) es siempre 1, haciendo que el peso correspondiente, w_0, sea una constante. Así se obtiene lo representado aquí como b_0 (Alpaydin, 2010, pp. 237-238). w_i y b_0 se pueden considerar como el conjunto de perillas de una red neuronal, las cuales se ajustan con muchísima precisión para dar la forma requerida al modelo.

Una vez obtenido el resultado (*z*) del algoritmo del perceptrón, este se alimenta a la función de activación mencionada más arriba. Existen múltiples funciones de activación con distintas utilidades que se utilizan comúnmente en las ANN. Se elige una de ellas dependiendo del fin con el que se haya desarrollado el modelo de red. Todas comparten la propiedad de limitar el *output* de los perceptrones individuales dentro de un rango controlado, independientemente del resultado de *z*, además del hecho de que la gran mayoría de ellas tienden a ser no lineales (a diferencia del algoritmo de un perceptrón individual). El valor obtenido a través de las funciones de activación, es el que termina estableciendo si el perceptrón se activa y funciona como *input* para nodos en capas

posteriores. Nótese que se trata de una versión de la ecuación (I). Es decir, los perceptrones funcionan con regresiones lineales simples. La función de los *outputs* con respecto a los *inputs* deja de ser lineal una vez que se introducen, tanto las funciones de activación, como las capas ocultas en el modelo, a pesar de que el algoritmo básico de cada neurona particular sea, fundamentalmente, una regresión lineal.

Con lo anterior se explican únicamente los perceptrones, que son los nodos (o "neuronas") de una ANN. Pero aún queda por aclararse cómo estos forman la estructura de la red. Hay varias arquitecturas de redes que se pueden utilizar como modelo (Graupe, 2013, p. 11), pero la que más se utiliza es la red perceptrón multicapa, que usa el algoritmo de propagación hacia atrás (*Back-Propagation network*) como única forma de aprendizaje (y que soluciona la incapacidad de modelar la o excluyente que se mencionó anteriormente. En esta arquitectura, la ANN está constituida por varias capas —conjuntos de neuronas—, entre las cuales se encuentran la capa de *inputs*, la(s) capa(s) oculta y la capa de *outputs*.

La primera está conformada por perceptrones que reciben los conjuntos de datos en su forma "cruda". Por ejemplo, en una ANN diseñada para distinguir si una imagen muestra un tiburón o no, la capa de *inputs* es el conjunto de nodos que toman sus valores de cada pixel (*input*) de las distintas imágenes. Así, se evidencia un primer nivel de complejidad en los modelos de ANN: las capas generalmente están conformadas por cuantiosísimos nodos. En el caso de la identificación de una imagen, suele haber un nodo por pixel. Cada uno de estos nodos, después de ser entrenados con muchísimas imágenes, se conectará mediante pesos distintos y muy específicos a la siguiente capa. Esta segunda es la capa oculta. Recibe los resultados de pasar por la función de activación el producto de los *inputs* y los pesos en las conexiones. A su vez, esos resultados se multiplican con los pesos de las conexiones con la siguiente capa, que puede ser otra capa oculta, o la capa de *outputs*. Todas las capas de nodos que se encuentren entre la de *inputs* y la de *outputs* son las que se denominan "capas ocultas". Una vez que han pasado por todas las capas ocultas —todos los pesos en las conexiones y las funciones de activación de los diferentes nodos—, llegan a la capa de *outputs*, donde el *input* completa su transformación algorítmica y se obtiene un resultado final.

La cantidad óptima de nodos por capa, así como la cantidad óptima de capas ocultas, es un tema que aún se discute (Stathakis, 2009). Sin embargo, lo que es cierto es que, si se utilizan más nodos de los necesarios en una capa, o más capas ocultas de las necesarias en una red, se corre el riesgo de sobreajustar el modelo. Las capas ocultas de una ANN cumplen el rol de otorgar mayor

complejidad a la red a través de una mayor cantidad de parámetros, los cuales incrementan la capacidad del modelo, entendida como la habilidad de ajustarse a una variedad más amplia de funciones (Goodfellow *et al*., 2016, p. 110). Un ejemplo de fines de gran complejidad para los que se requieren modelos de DL es el del campo de la biología conocido como genómica, el cual se dedica al estudio de conjuntos completos de ADN de distintas especies. La genómica produce cantidades masivas de datos. De hecho, se estima que para 2025 va a requerir de más espacio de almacenamiento que *Twitter*, *YouTube* y la astronomía combinados (Stephens *et al*., 2015, p. 2). Los datos que genera son, además, sumamente heterogéneos, lo que complica el análisis. En un caso de complejidad y cantidad tal, es donde se necesitan modelos de DL (Koumakis, 2020). Otros modelos menos complejos estarían lejos de poder ajustar sus funciones a los conjuntos de datos tan dispares y cuantiosos, para lo que se requieren muchísimos parámetros y flexibilidad en sus ajustes.

No obstante, aunque puede ser muy beneficioso en casos como el mencionado, el uso de ANN no puede ser recomendable de forma universal. El motivo no solo radica en los conjuntos de datos a partir de los que se quiere desarrollar un modelo, que, si es muy reducido, podría provocar un subajuste (*underfitting*), además de que, relativamente a un modelo más simple, la precisión se vería afectada por rendimientos decrecientes. También radica en el hecho de que las ANN requieren de muchos recursos computacionales. Si una tarea puede ser realizada por un modelo de regresión lineal con precisión similar a la de una red neuronal, ¿por qué se haría con la segunda? Al mismo tiempo, la interpretabilidad podría verse perjudicada innecesariamente a cambio de una insignificantemente mayor precisión. Una red neuronal de perceptrón multicapa puede tener el mismo nivel de precisión que un modelo de regresión lineal con un conjunto de datos reducido, pero se sacrificaría en parte su interpretabilidad, la capacidad de comprender cómo está funcionando el modelo, sin ningún beneficio real adicional más que una precisión marginalmente mayor.

Cabe aquí una aclaración sobre el concepto de “precisión”, el cual tiene varios énfasis y significados. Por un lado, puede hacer hincapié en los procesos de medición propiamente (psicometría, econometría, etc.). Por otro lado, y en la acepción que aquí se utiliza, la precisión puede hacer referencia a qué tan ajustado está un modelo al sistema que modela. Piénsese, por ejemplo, en un modelo simple de regresión lineal. En la medida en que modela la distribución de “felicidad reportada” con respecto a “ingresos anuales”, podría llegar a un alto nivel de precisión y,

por tanto, de tener predicciones correctas en una proporción mayor, ya que ambos fenómenos están correlacionados. Por otro lado, si lo que se intenta modelar es la velocidad de un fluido que corre a través de canales de otros fluidos complejos, entonces un modelo lineal no bastará y, de utilizarse, dará resultados imprecisos: resultados que no permitirían una proporción alta de predicciones correctas al respecto. Cuando se hable de precisión a lo largo de esta investigación, se hace referencia a este último sentido.

Para finalizar con la presentación de los modelos de ML que representan tres niveles distintos de complejidad, falta la explicación sintética de cómo transcurre el proceso de entrenamiento de las ANN. Para producir un modelo adecuado al fin que se tiene, se debe alimentar la red con datos de entrenamiento. Estos pueden variar en su nivel de estructura (con etiquetas o sin etiquetas), pero deben ser cuantiosos para obtener buenos resultados. Por ejemplo, si se quiere entrenar un modelo que reconozca entre un perro y un gato en una imagen que puede tener ambos, hay que brindarle muchísimas imágenes que incluyan perros, perros y gatos o ninguno de los dos, además de las respectivas etiquetas que señalen cuando hay uno, o los dos, o ninguno. Es más fácil que el resultado sea generalizable (y no se caiga en sobreajuste), si los perros y gatos se encuentran en distintas posiciones en las fotos, además de que sean de diversos tamaños, colores y formas.

Antes de ser entrenada, la red debe empezar con parámetros completamente aleatorios, para evitar parcialización en los resultados y para romper la simetría entre los diferentes nodos, especialmente los que se encuentran uno al lado del otro y están conectados al mismo *input*: si todos los nodos empezaran con pesos de 0, el modelo no cambiaría del todo (no se rompe la simetría). La aleatoriedad de los pesos iniciales se obtiene asignando algún número cercano a 0 (o 0) mediante la aplicación de un método de inicialización ofrecido por la biblioteca utilizada[20]. Una vez que se obtienen el primer *output* de un ejemplo de entrenamiento, se saca la diferencia de ellos con respecto a los *outputs deseados*, y se utiliza el cuadrado del resultado de esa diferencia para obtener el "costo" de esa diferencia (se utiliza el cuadrado pues limita los resultados a números positivos y, especialmente, porque le da más peso a rangos mayores de error que si tan solo se sumaran las variables no-cuadradas). Este es el método de los mínimos cuadrados que se mencionó anteriormente. Se repite lo mismo con todos los ejemplos de entrenamiento disponibles, y se

[20] Por ejemplo, en *Keras* se puede utilizar RandomNormal para generar esos pesos iniciales según una distribución normal. La distribución no es aleatoria, pero sí lo es la asignación de esos números o tensores a pesos específicos.

adquiere el promedio de costos, que da el "costo total" de la red. El modelo se ajusta más a los datos, cuanto menor sea el costo total de la red, es decir, cuando el rango de error entre el *output* obtenido y el *output* deseado se aproxime al menor posible. Por ello, una vez que se obtiene un primer costo total, se modifican los pesos y sesgos de toda la red, intentando reducir al máximo ese costo total.

Para alcanzar esta finalidad, se debe iterar en numerosas ocasiones con los mismos ejemplos de entrenamiento. Este procedimiento se realiza mediante el algoritmo conocido como *propagación hacia atrás* (mencionado anteriormente al hablar de la arquitectura de perceptrón multicapa), y funciona modificando primero los pesos y sesgos de la capa de neuronas más cercana a la de *outputs* para influenciar los resultados obtenidos en esta última. Después, modifica los pesos y sesgos de la siguiente capa, y así hasta llegar a la capa de *inputs* (Larochelle *et al*., 2009). Esta modificación se logra mediante el cálculo de la derivada del costo de un *output* específico (con respecto a *uno* de los *inputs* de entrenamiento) con respecto al peso de una conexión particular que produce ese *output*. Es decir, y por la naturaleza de las derivadas, se calcula cuán sensible es el costo frente a los cambios que se den en el peso. Cuanto más sensible sea, más eficiente es cambiar el valor de esos pesos que el de otros que sean menos sensibles para obtener el *output* deseado. Por la regla de la cadena, se puede representar ese proceso de la siguiente forma:

(III) $$\frac{\partial C}{\partial w} = \frac{\partial C}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w}$$

Donde *C* representa el costo del *output* obtenido con respecto al deseado, *w* representa el peso de la conexión específica, *y* representa el valor obtenido por el nodo tras modificar el *input* con el peso y sesgo correspondientes y pasar ese resultado por la función de activación y *z* representa ese mismo valor *antes* de pasarlo por la función de activación (Alpaydin, 2010, 249-250).

Este mismo proceso de cálculo de los pesos y sesgos más sensibles al cambio y su posterior modificación se aplica en cada conexión de la red, iniciando desde las capas más cercanas al *output*, y en dirección a las capas más cercanas al *input*, hasta obtener el modelo en el que el *output* deseado y el obtenido se aproximen, de tal forma que se optimice su utilidad. Esto es con respecto a un único ejemplar de entrenamiento, y se debe repetir para *cada ejemplar* de entrenamiento. Con lo anterior, se espera transmitir la complejidad de una red neuronal multicapa, y por qué podría el resultado de una de ellas en específico ser en principio inaccesible para la interpretación de un ser humano.

Habiendo aclarado, aunque superficialmente, el funcionamiento de los distintos modelos de ML, solo queda aclarar un último concepto antes de pasar a exponer con más detalle el problema de la opacidad epistémica. La siguiente sección explorará el significado del término *Big Data*, así como su relación con ML.

1.4 *Big Data*: cantidad y cualidad

Su nombre expresa, en realidad, solo parte de su naturaleza, pero deja por fuera muchísimos elementos que los hace ser tan relevantes en la actualidad. No es inexacto decir que BD se caracterizan por un salto cuantitativo muy apreciable en los datos recolectados y almacenados, pero sí es sumamente impreciso dejar la definición en ese punto. El concepto, no obstante, ha tenido cambios en su significado en tan solo un par de décadas. A pesar de su popularidad en los últimos años, BD como término se remonta a la década de 1990, donde hacía referencia a conjuntos de datos demasiado grandes para cualquier computadora de la época (Crawford *et al*., 2014). Era más un "problema" que una tecnología, una herramienta o un recurso, tal como aparece en Cox & Ellsworth (1997, p. 235), pero solo su nombre era nuevo. El problema data de inicios de la segunda mitad del siglo XX, cuando también existían dificultades "lógicas y lingüísticas, de *hardware* y de *software*, prácticas y teóricas" (Levien & Maron, 1967, p. 715) para lidiar con conjuntos muy amplios de datos. Este es el que se concibe como el problema epistemológico de BD, cuya solución, se cree, es simplemente tecnológica (Floridi, 2019, p. 103). Es decir, basta con mejorar el *hardware* y *software* que se utiliza para lidiar con esos conjuntos enormes de datos para resolver el problema. En el siguiente capítulo se verá que no es este el problema (Floridi, 2019, pp. 103-104), y que se deja muchísimo por fuera al plantear la solución como un asunto meramente tecnológico.

A pesar de lo anterior, el almacenamiento de datos, debido a su cantidad y peso, no deja de ser un problema del todo junto con la tecnología que ayuda a manejarlo: la Escuela de Información de Berkeley estimó hace unos años que, antes de la aparición de las computadoras personales, la humanidad había acumulado lo equivalente a un video de 50.000 años en calidad DVD, es decir, 10^{18} bytes (o 1 exabyte) (Floridi, 2014). Un informe de la *International Data Corporation* (IDC) del 2012, calculaba que en el 2020 se alcanzaría un total de datos que rondaría los 40.000 exabytes. Esto equivale a 40.000 veces lo que la humanidad había producido en la totalidad de la historia previa a las computadoras (Gantz & Reinsel, 2012). Al mismo tiempo, la IDC también calculó que, para el

2020, la capacidad de almacenamiento disponible globalmente sería de 6.8 zettabytes (6.800 exabytes) (Reinsel *et al*., 2018), lo cual no es ni un cuarto de la cantidad total de datos producidos. Evidentemente, no se busca almacenar la totalidad de los datos producidos, pero el asunto es que siempre se está en un déficit de almacenamiento en términos generales, lo cual representa una limitación tecnológica que, aunque no se esté dando necesariamente en la práctica, teóricamente implica problemas potenciales a futuro (cercano).

A pesar de lo anterior, BD ha dejado de hacer referencia a un problema como tal, aunque el problema al que hacía referencia el término hace unas décadas aún persiste. El poseer una cantidad tan abundante de datos y poder manejarlos más o menos efectivamente resulta más bien en un neto positivo antes que un problema, al menos en términos de producción de conocimiento (Floridi, 2019, p.103). De esta forma, BD parece hacer referencia actualmente más bien al *uso* que se da de esos cuantiosísimos datos, así como al conjunto de tecnologías y prácticas que permiten aprehenderlos.

Como su nombre indica, el mero tamaño de estos conjuntos de datos es una de las características salientes de BD. Es preciso ejemplificarlo para tener una idea de sus dimensiones. Walmart, por ejemplo, con 20.000 tiendas a lo largo del mundo, generaba en 2016 2.5 petabytes de datos por hora (Marr, 2016). Esto equivale a 2500 terabytes (1 terabyte=10^{15} bytes) (se volverá al caso de Walmart posteriormente, ya que ejemplifica de forma muy detallada las diversas características que hacen de BD un fenómeno capaz de provocar un nuevo paradigma en la producción de conocimiento). Alternativamente, Facebook tiene 1730 millones de usuarios activos cada día (Facebook, 2020), por los que genera 4 petabytes (4000TB) diarios de datos (Wiener & Bronson, 2014). La cantidad de información que amasan estas compañías y similares, requiere de tecnología especial para poder ser procesada y analizada efectivamente. Una única computadora, aunque sea la más poderosa de la actualidad, no estaría ni cerca de ser capaz de lograr esta tarea. Por ejemplo, una de las herramientas más utilizadas para el análisis de datos en empresas, las hojas de cálculo de Excel, tiene un límite de 2 gigabytes para su espacio de direcciones virtuales (*Excel Specifications and Limits*, s/f). En términos de filas, Excel tiene un límite de 1.048.576 posibles en una sola hoja de cálculo. Walmart, por su parte, tiene un centro de datos con 200.000.000.000 filas compuestas de información transaccional de los clientes (Marr, 2017). Evidentemente, Excel se queda corto por sí mismo.

Parece claro que la mera cantidad de datos que manejan ciertos entes (instituciones, corporaciones, empresas, etc.) es fundamental para comprender lo que hace a los BD. Sin embargo, el análisis no puede quedar solo en esta característica. Lo que ocurre es que la cantidad está acompañada de muchos otros elementos que le dan a BD su utilidad e importancia. En otras palabras, BD no se debería definir únicamente como colecciones enormes de datos, las cuales escapan a la capacidad de computadores individuales. Esta definición sería muy simplista, y no capturaría lo que se podría considerar la esencia de este fenómeno.

Con esto en mente, en los últimos años se han realizado investigaciones sobre la literatura dedicada al tema, con el fin de extraer de ella una definición pragmática (obtenida de cómo se utiliza realmente en el ámbito académico) y más precisa (De Mauro *et al*., 2015 y 2016). En algunos casos, la metodología consiste en la revisión de artículos y libros, tanto de la academia como de la industria. De hecho, el fenómeno no apareció primero en círculos académicos, sino en el sector empresarial privado. Por lo general, las nuevas tecnologías o teorías pasan por un período de análisis académico y técnico antes de que su uso se popularice en las distintas industrias y en el sector empresarial privado. En este caso, el uso vino antes del análisis, por lo que el acercamiento académico apenas está empezando a madurar (Gandomi & Haider, 2015) y no obstante, cuantiosos artículos y conferencias se han presentado sobre el tema.

De Mauro *et al.* (2016) extrajeron una lista de 1581 piezas escritas en las que el término *Big Data* estaba incluido en el título o en las palabras clave. Con base en este conjunto de datos, hicieron un análisis que los llevó a determinar los conceptos más frecuentemente utilizados al lado de BD. Encontraron que comúnmente esos conceptos hacen referencia a cuatro elementos relacionados con BD: (1) información, (2) tecnología, (3) impacto y (4) métodos. Cada uno de ellos funciona como un conjunto con elementos que, en ocasiones, se comparten con otro de los conjuntos. Por ejemplo, al hablar de información, se habla de la distinción entre información estructurada y no estructurada, se habla de la "datificación" (*datafication*) y de la privacidad de la información. Los dos últimos, pertenecen también al conjunto 3, ya que principalmente tiene elementos que se refieren a la influencia de BD en la sociedad. El conjunto de tecnología tiene elementos como la capacidad de almacenamiento, la computación en paralelo y demás, mientras que el de métodos incluye algunos de los modelos que se han revisado en esta investigación, de ML, DL y demás. (De Mauro *et al*., 2016, p. 128).

Sumado a lo anterior, hicieron un análisis de las definiciones explícitas que se ofrecen del concepto en esa literatura. Encontraron que las definiciones responden a otros cuatro grupos categoriales distintos: definiciones que enfatizan (1) atributos de los datos, (2) necesidades tecnológicas, (3) umbrales e (4) impacto social. En 1 se remarcan características como la velocidad, el volumen, la variedad, la complejidad y falta de estructura de los grandes conjuntos de datos. En 2 se habla esencialmente de la capacidad tecnológica requerida para el procesamiento de esos conjuntos, como la capacidad de almacenamiento y manipulación que va siempre en crecimiento. Para 3, se expone BD como algo que depende de cruzar algún umbral: un conjunto de datos se considera BD cuando, por ejemplo, se requiere más que la capacidad de una computadora individual (de alta gama) para poder almacenarlos y manipularlos adecuadamente. Finalmente, en 4 se reúnen las definiciones que recalcan la influencia social que pueden tener, por ejemplo, definiendo BD como un fenómeno cultural, tecnológico y académico que cambia nuestro entendimiento de la sociedad (De Mauro *et al*., 2016, p. 130).

Con base en el análisis, los autores concluyen con una definición formal del concepto de BD:

> *Big Data* es el recurso de información caracterizado por volumen, velocidad y variedad tan altos, que requiere de tecnologías y métodos de análisis específicos para su transformación en algo de valor. (De Mauro *et al*., 2016, p. 131).

La definición es útil en la medida en que incluye varios de esos aspectos que los autores encontraron como mayoritariamente ligados al concepto: atributos de datos, tecnologías e impacto social. Queda por fuera el concepto de "umbrales", y se puede inferir que se debe a que definir un término según umbrales puede generar confusiones, siendo que la elección del umbral no podría no ser arbitraria ulteriormente. En otras palabras, establecer un umbral solo puede realizarse por convención y basado, probablemente, en capacidades computacionales promedio, o alguna cosa semejante.

En adelante, cuando se hable de BD en esta investigación, se hará referencia a la definición que provee el análisis detallado de estos autores. En especial porque permite establecer con claridad la relación entre estos y los modelos de ML más complejos. Se hablaba en secciones anteriores que, cuanto más complejo era el conjunto de datos por analizar, tanto más complejas tendían a ser las herramientas tecnológicas que permitían ese análisis. Por lo general, cuando se habla de procesar y analizar BD, se habla también de métodos con modelos de ANN, especialmente las que pertenecen al subconjunto de DL (Najafabadi *et al*., 2015).

Más aún, muchas veces se requiere de métodos sin supervición de DL, debido a que la gran mayoría de datos que se recolectan en la actualidad, incluso los que son específicos a ciertos dominios, son datos no estructurados y, por lo tanto, sin etiquetas (Dhar, 2013, p. 66). No existe mano de obra humana que pueda etiquetar eficiente y eficazmente esos datos para poder entrenar modelos de forma supervisada, por lo que se requiere de redes neuronales que no necesiten de ella para transformar esos datos, ahora sí, en información etiquetada. Muchos de estos datos, aunque no todos, son incluso subsimbólicos, lo cual dificulta aún más su clasificación o análisis de grupos. Estas, como se podría esperar, implican un crecimiento en la complejidad y, por lo tanto, una disminución de su interpretabilidad. Por esto último es que es importante en esta investigación la combinación de BD y ML, porque es en estos casos donde más sobresale la opacidad epistémica cuya naturaleza se va a introducir en la siguiente sección. Se han presentado los conceptos básicos de la investigación sobre los que se podrán formular argumentos en los siguientes capítulos: se definió lo que se entiende por *Big Data*, así como ML, DL y algunos modelos específicos que muestran el rango entre los más simples y los más complejos. En el siguiente capítulo se analizará a fondo el problema que acompaña a esa complejidad: el de la opacidad epistémica. Se explorará un campo de la ciencia de datos que pretende eliminar o, al menos, mitigar ese problema: la interpretabilidad de modelos. Con esto, se profundizará mucho más en la naturaleza de esa opacidad, en las técnicas que se han diseñado (y que se están diseñando) para esquivarla o, en el mejor de los casos, superarla, y en su aplicación para cuestiones relacionadas con las ciencias naturales. Para ello, se propone en esta investigación dividir el problema de la caja negra en dos partes: el problema desde su perspectiva técnica, y el problema desde su perspectiva epistémica. En el siguiente capítulo se comenzará por aclarar esta distinción de manera detallada.

Capítulo 2: Cajas negras e interpretabilidad de modelos: una aproximación epistemológica

El capítulo anterior explicitó el área temática de esta investigación. Se presentaron los tres elementos constituyentes de la ciencia de datos: la estadística matemática, las ciencias computacionales y los dominios particulares en las que se aplica. Se entendió la ciencia de datos como una continuación de la estadística en tanto estadística aplicada. También, se aclaró el subconjunto de la IA conocido como *machine learning*, y se revisaron algunos de sus conceptos principales, tales como la estructuración o no-estructuración de los datos, la IA simbólica o subsimbólica, el sobreajuste y la flexibilidad de los modelos, así como los beneficios que provee el ML a la producción de conocimiento. Posteriormente, se profundizó sobre las distintas aproximaciones metodológicas de ML que se suelen utilizar, yendo de las menos complejas a las más: regresión lineal, árboles de decisión y, finalmente, redes neuronales multicapa. La revisión de tales estrategias facilita la comprensión del nivel de complejidad involucrado en modelos de *Deep learning*. Finalmente, se dio una definición técnica de *Big Data*, término que se utiliza coloquialmente, pero no está tan claramente determinado en la literatura especializada, y se mostró su particularidad frente a conjuntos de datos más fáciles de manejar y por qué se pueden entender como un fenómeno distinto, no solo por cuestiones cuantitativas.

Este capítulo tiene un propósito doble: (1) exponer el problema de la opacidad epistémica en un subconjunto de modelos de ML, también conocido como el problema de la caja negra (*black-box problem*) —a lo que estará destinada la sección 2.1—, y (2) mostrar los avances, defectos y limitaciones de la inteligencia artificial explicable y la interpretabilidad de modelos, áreas de la ciencia de datos que plantean métodos y teorías para hacer comprensibles los modelos de ML, ya sean más o menos opacos —lo cual corresponderá a las secciones 2.2, 2.3 y 2.4—. Se hará un análisis de las conclusiones y preguntas a las que ha llegado la literatura actual sobre el tema, las discusiones del cual están muy activas presentemente, para finalmente ofrecer una mirada epistemológica que aclare conceptos usualmente utilizados y malentendidos.

2.1 La opacidad epistémica en ML: El problema de la caja negra

Aclarados los conceptos relevantes para comprender este problema en las secciones anteriores, nos es posible presentar a continuación el problema central que concierne a esta investigación. El problema de la opacidad epistémica que producen los modelos complejos de ML, especialmente aquellos que se modelan sobre la base de conjuntos de datos que podrían llamarse BD, se suele conocer en la literatura también como el "problema de la caja negra" (*the black box problem*) (Rudin, 2019; Lipton, 2018; Brendel *et al*., 2018; Papernot *et al*., 2017; Krause *et al*., 2016; Watson *et al*., 2019; Castelvecchi, 2016; Koh & Liang, 2017, etc.). Se hace referencia a una caja negra cuando no es posible responder a la pregunta de por qué un modelo obtiene los *outputs* que obtiene. Estos *outputs* pueden ser los correctos, en el sentido de que eran los que se esperaban obtener y son efectivos al aplicarse, pero siempre queda la falta de comprensión sobre cómo se llegó del conjunto de *inputs* inicial, a los *outputs* finales. La relación entre BD y el problema de la opacidad, es que aquellos, por su cantidad, suelen requerir de técnicas de ML complejo para su agrupación, etiquetación o análisis en general.

Aquí hay una distinción importante. Se habla de comprensión de modelos de ML complejos en dos sentidos: (1) se comprende cómo funcionan *los algoritmos* en general, o cómo funciona *un algoritmo* en específico, puesto que se comprenden las líneas de código que se utilizan para programarlo y se pueden explicar los procesos utilizados para su entrenamiento. De hecho, las líneas de código requeridas para programar una red neuronal relativamente compleja no son demasiado complejas en sí mismas. En otras palabras, se comprenden *las reglas* mediante las que aprende y se forma el modelo (Lillicrap *et al*., 2019, p. 1). Aquí podríamos hablar, incluso, de que se comprenden los detalles en un sentido computacional o algorítmico, pero no se comprende el nivel *semántico* de estas herramientas. La parte *semántica* hace referencia a que (2) no se comprende cómo funciona el modelo una vez ha sido entrenado; no se comprende en muchos casos por qué da énfasis a ciertas características sobre otras, por qué los pesos y sesgos con los que se optimiza son los adecuados para, por ejemplo, determinar que una imagen es un gato y no un caballo, etc. Llamemos a (1) la comprensión *técnica* o computacional y a (2) la comprensión *epistémica*[21] o semántica.

La falta de comprensión *técnica* puede darse por varios motivos. Por ejemplo, alguien que no tenga la pericia necesaria para comprender cómo funcionan ciertos modelos de ML, podría ver

[21] Haciendo referencia a las virtudes intelectivas que presenta Aristóteles en su *Ética Nicomáquea*.

como una caja negra incluso a los modelos más simples de regresión lineal o de árboles de decisiones. Esta percepción es importante, dado que afecta la confianza que los usuarios no expertos tienen sobre el funcionamiento de la inteligencia artificial para fines útiles. Sin embargo, esta clase de caja negra es accidental, dado que se puede relegar únicamente a la falta de pericia de quien percibe el modelo como opaco, más que al modelo mismo. En este mismo sentido, las ciencias naturales podrían llamarse una caja negra para los que no son expertos en ellas. La razón por la que no se puede sobrepasar la velocidad de la luz, por ejemplo, es completamente opaca para quien no maneje los conceptos básicos de la teoría de la relatividad especial.

Otra caja negra accidental que se puede atribuir a un tipo de falta de comprensión *técnica* es la que ocurre cuando los modelos son opacos, no por su complejidad, sino porque son propiedad privada de algún ente, ya sea instituciones, empresas o individuos. Son cajas negras porque es imposible *legalmente* conocer su funcionamiento (Rudin, 2019, p. 206). Se puede pensar en ellos como la "receta secreta" de algún producto en el menú de un restaurante. Se mantiene oculta porque, si se pudiera replicar, no podrían obtenerse ganancias económicas del modelo. Este es el caso de COMPAS, un modelo privado de reincidencia en actos criminales utilizado por juzgados de diversos estados en Estados Unidos. En realidad, el modelo es bastante simple, ya que depende de pocas características (variables) y es probablemente lineal (Rudin, 2019, p. 208), por lo que, si fuera público, no sería ninguna clase de caja negra. Pero su condición de propiedad privada lo hace completamente opaco a la inspección pública. De esta forma, sus dueños generan ingresos, ya que ofrecen una herramienta con la precisión que desean muchos de los juzgados a los que sirve (fuera de la controversia posterior con el análisis de ProPública (Rudin, 2019, p. 208)). Pero, como en el caso anterior, se trata de una opacidad no adjudicable a la naturaleza del modelo, sino a una condición externa accidental.

Estos ejemplos se distinguen de la falta de comprensión *epistémica*. Como se ha insinuado anteriormente, cuando un modelo produce una función demasiado compleja para el entendimiento humano (Rudin, 2019, p. 206), tal que no se comprenda adecuadamente la causalidad que hace que un *input* (x) termine convirtiéndose en un *output* (y), entonces se puede denominar a ese modelo una caja negra, y no ya una accidental. Este caso difiere del del individuo con falta de pericia, en el sentido de que él podría, en principio, comprender la función del modelo —siempre que se dedicara a aprender sobre ML, estadística y el dominio al que se aplique—, mientras que, en aquel, la

opacidad es inevitable, dado que la complejidad del modelo obtenido es tal, que la capacidad de un ser humano para comprenderlo sería solo superficial, nunca completa, ni de cerca.

¿Qué genera tal nivel de complejidad? Para responder, es importante hacer uso del conjunto de conceptos explicados más arriba relacionados con redes neuronales artificiales, *Deep Learning* y *Big Data*. Una red neuronal compleja puede variar en su constitución en la cantidad de nodos por capa, y en la cantidad de capas ocultas en la red. Cuanto mayores sean esas cantidades, tanto más difícil de comprender el funcionamiento del modelo, ya que introducen una mayor cantidad de pesos y sesgos, así como más y distintas relaciones entre ellos. Hay que recordar que, en una arquitectura convencional de ANN, cada nodo de una capa está conectado directamente con *cada nodo* de la capa que le sigue. Es decir, las interacciones entre las neuronas en diferentes capas son demasiado variadas y cuantiosas para poder seguirlas.

A diferencia de un árbol de decisión, en el que se habla de una única dirección de división recursiva de los datos ("de arriba hacia abajo"), lo cual hace al proceso en principio fácil de seguir, en una ANN con miles de nodos y varias capas ocultas, la única dirección de la que se puede hablar es de *input* a *output*. Y esta es una dirección tan general y poco significativa, como decir que, para llegar a cierto lugar, uno debe caminar del punto de partida al punto de llegada. Los diferentes pesos y sesgos en los nodos son tantos, tan específicos y tan variados, que la influencia que tiene cada nodo sobre el siguiente es muy difícil de determinar, ya que ella depende, a su vez, de los nodos que los preceden (con sus respectivos pesos y sesgos). Esta complejidad es exactamente lo que permite a estos modelos encontrar patrones que para un ser humano serían imposibles de detectar, y se debe a la granularidad que permite la existencia e interconexión de tantos parámetros. Pero es ella misma la que repercute en la capacidad de un individuo de comprender exactamente lo que ocurre dentro de la red.

Se puede ver este mismo punto desde otra perspectiva para mayor claridad: En los nodos de salida (*output*) existe un nivel semántico comprensible por lo general. En los nodos de entrada (*input*) podría haberlo. Pero en las capas intermedias de la red neuronal, no habría ningún tipo de sentido semántico. Se trata de una versión desestructurada de la entrada y de lo que después será la salida, la cual, por esa misma desestructura, no se puede considerar que haya sentido en cada nodo.[22] Sobre BD desde esta perspectiva, la cantidad de datos no implican nada sobre su nivel

[22] Esto es interesante, porque hace referencia de alguna manera al atomismo lógico de Wittgenstein (2012) y Russell (2009). Habría ciertos mínimos de sentido que no podrían ser reducidos en mayor medida.

semántico, pero sí cuando se habla sobre las relaciones o las formas de agrupar esas cantidades de datos. Las variables por sí mismas pueden o no tener semántica, pero encontrar relaciones entre esa totalidad es lo que no produce semántica.

Ahora, combínese lo expuesto con conjuntos de datos tan masivos, que la sociedad humana tiene problemas de espacio para almacenarlos, incluso con tecnologías recientes que buscan mitigar este problema técnico. Datos que, como se vio arriba, son cada vez más, y cada vez menos estructurados. Se había visto en la sección 1.2 que el etiquetar los datos, así como la selección de sus características centrales, aporta mucha interpretabilidad a los modelos. Pero, no solo son las *deep neural networks* (DNN) extremadamente complejas por la estructura e interacción de sus parámetros, sino que los datos con las que muchas de ellas se entrenan (sin supervisión) y sobre los que se espera obtener resultados efectivos y precisos, son carentes de estructura en su gran mayoría. En razón de ello es que se elige pensar en esta combinación de elementos, en la que la interpretabilidad del modelo se puede colocar en el nivel más bajo, al menos según la capacidad tecnológica explicativa e interpretativa que se tiene en la actualidad.

La interpretabilidad no es problema en muchos casos, como cuando se busca optimizar únicamente la precisión de los modelos. Así se comportan muchas empresas privadas, las cuales utilizan BD y ML para producir varios efectos deseados que, finalmente, repercuten en aumentos de su capital. La interpretabilidad de los modelos que aplican podría ser beneficiosa, claro, pero no es *requerida*. Pero si volvemos el enfoque hacia las ciencias, el problema se vuelve mucho más relevante. Asumiendo que la finalidad ulterior de la ciencia es la comprensión precisa del mundo en sus diferentes dimensiones, empieza a asomarse el problema que puede generar la aplicación de estas “cajas negras”. ML, combinado con BD, puede dar resultados sumamente precisos sobre fenómenos cuyo estudio hace unos cuantos años no estaba al alcance de los humanos del todo. Sin embargo, para dar esos resultados precisos, se debe pagar con cierto nivel de opacidad epistémica, la cual juega en contra de la esencia misma de las ciencias, como se profundizará con más detalle en el capítulo 3.

De manera más precisa, la opacidad epistémica se da cuando un agente cognitivo X en el tiempo t no conoce todos los elementos epistémicamente relevantes de un proceso (Humphreys, 2009, p. 618). Es decir, la opacidad no pertenece al proceso o al modelo en sí mismo, sino que pertenece a la relación entre ese modelo y un (cualquier) agente cognitivo. La relevancia epistémica se determinaría dependiendo del nivel de abstracción en que se requiera y, en consecuencia, de la

finalidad que se busque. Enfocándose en el agente cognitivo, la definición anterior de opacidad epistémica sería no esencial. La opacidad es esencial cuando el agente cognitivo X no conoce todos los elementos relevantes porque le es imposible por naturaleza (Humphreys, 2009, p. 618). Lo interesante de esta definición es que la relevancia epistémica solo se podría determinar *a posteriori* en casos no esenciales, pero en los esenciales se podría asumir *a priori* la imposibilidad de determinarla. Como aclaran además Kaminski y Schneider (publicación pendiente), la opacidad epistémica es opacidad metodológica, no opacidad de los objetos (los modelos o la naturaleza) en sí mismos.

Cabe preguntar, dadas las aclaraciones anteriores, si la opacidad epistémica de los modelos de ML complejo, o las simulaciones computacionales de sistemas físicos complejos, tiene una naturaleza tal que la hace diferente de la opacidad epistémica que parece ser inherente a los procesos científicos tal como se entienden hoy en día. Por ejemplo, la práctica científica no es, en la mayoría de los casos, una práctica individual. Es más bien una práctica gregaria por definición, dado que se requiere de revisión de pares, de trabajo conjunto de equipos, donde cada miembro cumple una función distinta y especializada, y dado que se depende constantemente de las conclusiones alcanzadas por otros de esos equipos como punto de partida para alcanzar nuevas conclusiones. Desde esta perspectiva, la ciencia como un todo es epistémicamente opaca para cualquier individuo, incluso uno experto en algún dominio científico: puede contribuir desde su especialización y el conjunto de tareas que le corresponden, pero no alcanza para conocer cómo llegan todos sus colegas a todas las conclusiones a las que llegan y los detalles de sus procesos. Tampoco alcanza para conocer absolutamente todas las herramientas tecnológicas (incluyendo teoremas y fórmulas matemáticas cuya prueba no es conocimiento requerido para su aplicación) y la razón por la que funcionan (Kaminski & Schneider, publicación pendiente).

¿Qué es diferente en la aplicación de modelos complejos de ML en la ciencia que da a su opacidad epistémica un lugar distinto al de la opacidad epistémica que ya es común en la práctica científica? Para responder a esta pregunta, Kaminski y Schneider (publicación pendiente) proponen tres tipos aspectualmente distintos de opacidad epistémica: opacidad social, opacidad técnica y opacidad matemática. La primera tiene que ver con la especialización de los investigadores y su subsecuente cooperación, la segunda se refiere a la aplicación de herramientas para la observación, la medición o la manipulación de objetos y la tercera se relaciona con el uso de herramientas matemáticas que, a pesar de ser efectivas, no se comprenden completamente a lo interno. Estas

formas de opacidad son *aspectualmente* distintas, no *clasificatoriamente* distintas, en el sentido de que su distinción no las reduce a diferentes clases de un fenómeno, sino que considera al fenómeno desde diferentes aspectos. Esto quiere decir que los distintos tipos de opacidad se relacionan entre sí: la opacidad tecnológica es un tipo de opacidad social, así como la opacidad matemática podría ser considerada un tipo de opacidad tecnológica.

A esto se suma una consideración más: la opacidad matemática no se da en el clásico caso en que se utiliza algún teorema con una función que cumple a cabalidad, pero cuya prueba no es conocida por la usuaria. Piénsese, por ejemplo, en una científica que hace uso de la fórmula de la fuerza gravitacional para determinar alguna cantidad que necesita. Sin embargo, no recuerda cuál es la prueba de la fórmula de la fuerza gravitacional: ¿por qué se multiplican las masas? ¿Por qué se utiliza el cuadrado de la distancia entre los centros de las masas como denominador? ¿De dónde se obtiene la constante gravitacional? Para sus objetivos, responder a estas preguntas no es necesario. Lo único que se requiere es saber que la fórmula funciona, y aplicarla tal y como se expresa. Incluso puede no conocerla por mero olvido, si conocía su prueba en el pasado[23]. No obstante, esta prueba es *en principio* derivable. Es decir, si la científica quisiera, podría estudiar sobre la fuerza gravitacional más a fondo y reencontrar y comprender a cabalidad la prueba que la verifica. Esto sería más cercano a la opacidad de tipo tecnológico, entendiendo la tecnología como lo hace Husserl: no algo necesariamente material, sino una técnica o herramienta que permite llegar a resultados sin la necesidad conocer por qué funciona como funciona, sino solo que funciona (*know-how* en lugar de *know-that*). Este no es el caso cuando se aplican modelos complejos de ML (Kaminski & Schneider, publicación pendiente)[24]. En el caso anterior se conoce a lo interno o se *puede* conocer a lo interno el funcionamiento de la tecnología específica, sea matemática o de otro tipo y a pesar de que no sea relevante para su uso. Los modelos complejos de ML tienen una estructura matemática que no se puede conocer completamente a lo interno. Ahí radica la diferencia.

[23] Lo cual contaría como otra clase de opacidad tecnológica: su versión del pasado que conocía cómo se probaba la fórmula fugiría aquí como el colega que conoce por qué funciona como funciona el instrumento tecnológico.

[24] En el artículo aquí referido, los autores utilizan estos argumentos para hablar sobre la opacidad epistémica de la simulación computacional en la ciencia. Se incluye aquí *ad hoc* a los modelos complejos de ML porque satisfacen criterios muy similares (si no exactamente los mismos) que la simulación, por lo que el argumento se adecúa completamente.

Como se dijo anteriormente, es posible tener una *comprensión técnica* de estos modelos: cómo funcionan sus algoritmos, las justificaciones estadístico-matemáticas por las que se utilizan y por qué permiten obtener resultados. Pero comprender, una vez entrenado el modelo, por qué específicamente se configura con los pesos y sesgos que lo hace, es *en principio* imposible, lo cual produce la opacidad aquí referida. Esa configuración, en un modelo suficientemente complejo, no es *en principio* derivable, como sí lo era la prueba de la fórmula de la fuerza gravitacional. No es derivable por los límites de la capacidad cognitiva humana, no por la naturaleza misma del modelo. Esta diferencia permite entender cómo la opacidad epistémica de los modelos de ML complejos es un fenómeno nuevo y distinto de la opacidad epistémica ya existente en la práctica científica en general.

Con lo determinado en los párrafos anteriores sobre la opacidad epistémica, también es posible ahora delimitar lo que implicaría, en contraste, la transparencia epistémica, la cual, como se verá en la sección siguiente, es la finalidad de la rama de la ciencia de datos conocida como *machine learning* interpretable. Si la opacidad epistémica se da cuando un agente cognitivo X no conoce todos los elementos epistémicamente relevantes en un tiempo t de un proceso, entonces la *transparencia epistémica* se daría cuando un agente cognitivo X *conoce todos* los elementos epistémicamente relevantes (con respecto a un nivel de abstracción y finalidad específicos) del proceso. Nótese que no se trata de comprender el proceso *simpliciter*, sino los elementos que son relevantes para el fin específico que se busca con él. No está de más reforzar la idea de que lo que se pretende no es comprender los modelos complejos de ML *completamente*, sino hacerlos transparentes en este sentido. Esta distinción es importante, por ejemplo, en discusiones sobre lo que implica un derecho a la explicación para cuerpos internacionales en torno al uso de ML para tomar decisiones sobre individuos, ya que algunos alegan doble estándar de parte de quienes piden ese derecho porque no buscan, al mismo tiempo, comprender *completamente* el aparato cognitivo humano (Zerilli *et al*., 2019). El asunto es que no es necesario comprender *completamente* el funcionamiento del modelo complejo de ML, sino únicamente los elementos relevantes.

Se ha presentado el problema de la opacidad epistémica relacionada con ML y BD y se ha distinguido entre la comprensión técnica y la comprensión epistémica. También se ha mostrado superficialmente su relación con las ciencias naturales. En la siguiente sección se explorará un campo de la ciencia de datos que pretende eliminar o, al menos, mitigar ese problema: la interpretabilidad de modelos. Con esto, se profundizará mucho más en la naturaleza de esa

opacidad, en las técnicas que se han diseñado (y que se están diseñando) para mitigarla o, en el mejor de los casos, superarla, y en su aplicación para cuestiones relacionadas con las ciencias naturales. Para ello, se propone en esta investigación dividir el problema de la caja negra en dos partes que responden a los dos sentidos de "comprensión" dados: el problema desde su perspectiva técnica, y el problema desde su perspectiva epistémica. En la siguiente sección se comenzará por profundizar sobre la primera, la perspectiva técnica.

2.2 Interpretabilidad de modelos: cómo abrir la caja

El problema de la caja negra no es nuevo, pero la creciente utilidad y aplicación de modelos de ML para la toma de decisiones relevantes, tanto para sociedades enteras como para los individuos que las constituyen, hace urgente la concentración de más esfuerzos por aclararlo y mitigarlo. Tal urgencia se ha visto reflejada en la cantidad de trabajo investigativo sobre el tema de interpretabilidad de modelos que ha surgido en años recientes (Doshi-Velez & Kim, 2017, p. 2).

La interpretabilidad de modelos o *Machine Learning* interpretable (IML por sus siglas en inglés, en adelante), es un área de la ciencia de datos que busca encontrar conocimiento relevante *en* un modelo (no necesariamente *mediante* el modelo) de ML respecto de las relaciones entre los datos que aprendió el modelo (Murdoch *et al*., 2019, pp. 22071-22072). En otras palabras, se estudian posibles métodos para hacer comprensibles los modelos de ML para los seres humanos en diferentes niveles epistémicos. Se habla de diferentes niveles epistémicos, pues las maneras de interpretar los modelos están sujetas al nivel de familiaridad técnica que se tenga con ellos. No se puede esperar la misma interpretación del modelo de parte de una bióloga experta en genómica que ha pasado años de su carrera utilizando redes neuronales para el análisis y la secuenciación de genomas, que de parte de una persona que no ha tenido mayor relación con la programación. El último caso se relaciona con la comprensión técnica presentada en la sección pasada.

Estas diferentes formas de interpretación se abordan de maneras muy distintas y respondiendo a diversas finalidades específicas, las cuales se pueden dividir taxonómicamente para obtener un panorama más claro. Una primera división (1) es la de (1a) interpretabilidad *inherente* frente a la (1b) interpretabilidad *post hoc* (Molnar, 2022, p. 21). Los modelos pueden ser desarrollados como interpretables por sí mismos desde el principio —interpretabilidad inherente—, aunque para hacerlo la clase de modelo de ML que se utiliza se limita mucho más, dado que solo

algunos pueden ser en principio interpretables por sí mismos. De hecho, algunos autores encuentran que hay una responsabilidad moral en utilizar modelos de suyo interpretables para aplicaciones que puedan afectar a la sociedad o sus individuos (Rudin, 2019).

Por otra parte, también es posible diseñar modelos de ML cuya finalidad sea *explicar* el funcionamiento de *otros* modelos que podrían considerarse cajas negras —interpretabilidad *post hoc*—. Este suele ser el paradigma utilizado para la explicación de modelos de suyo complejos, tales como los DNN. A fin de cuentas, la interpretabilidad inherente funciona en particular cuando la falta de comprensión es *técnica*, pero se queda corta cuando se trata de la *epistémica*, ya que en muchos casos esta responde a una falta de interpretabilidad en todo el proceso o en parte del mismo. O, en otras palabras, si un modelo no necesita de una explicación externa para poder ser interpretado, entonces es en sí mismo epistémicamente comprensible. Por este motivo, hablar de interpretación intrínseca como solución de la falta de comprensión epistémica de un modelo complejo es contradictorio.

No obstante, mientras que ambas formas de interpretabilidad son importantes y deben ser desarrolladas y estudiadas, en la actualidad el uso de interpretabilidad inherente podría ser en general más eficiente y efectivo. Los modelos más complejos de ML suelen ser sobreutilizados relativos al nivel de precisión y confianza que pueden producir hoy en día (Rudin, 2019), y muchas de las tareas en que se aplican podrían ser realizadas con significativamente mejores resultados mediante modelos más simples y, por lo tanto, más inherentemente interpretables de ML. Es importante recordar que la interpretabilidad no es deseable únicamente por sí misma, o para que los usuarios no expertos logren comprender los resultados de un algoritmo. También es necesaria para poder depurarlos de forma más adecuada, de lo cual se obtienen, a su vez, modelos más precisos y efectivos.

Existe en la literatura especializada una idea de que hay un intercambio de proporcionalidad inversa entre la interpretabilidad de un modelo y su precisión: cuanto más preciso sea, tanto menos interpretable se convierte. Esto podría teóricamente argumentarse como cierto, pero con algunas aclaraciones. Como se mencionó anteriormente, hay consenso con respecto al hecho de que las ANN son capaces de modelar *cualquier* función imaginable (Hornik et al., 1989). Esto implica lógicamente que con una ANN se puede idealmente obtener precisión máxima. En la práctica es más complejo por los recursos computacionales que consumiría una red con ese nivel de precisión, pero lógicamente es posible. Sin embargo, para que una función sea máximamente precisa, su nivel

de complejidad va a ser necesariamente mayor: cuanto más detalle se deba modelar, tanto más y más variadas especificaciones tendrá la red. Así, el intercambio de proporcionalidad inversa comentado es lógicamente posible. No obstante, en la práctica no necesariamente debe ser este el caso. Podríamos decir que existe un nivel de simpleza de funciones que modelan sistemas en las que convergen la interpretabilidad y la precisión. Conforme se va complejizando el sistema por modelar, la precisión empieza a bajar para modelos interpretables y a subir para modelos complejos. Hay un amplio espectro, sin embargo, en el que los modelos interpretables aproximan su precisión a la del modelo complejo, incluso si idealizamos la precisión máxima de este último. Es a este espectro al que se refiere Rudin (2019) cuando propone la ausencia de necesidad de trocar interpretabilidad por precisión, y es ese espectro el que suele ser ignorado en favor de precisión, como se verá más adelante en los problemas de utilizar modelos complejos en la ciencia.

Una nota sobre el término de "interpretabilidad". Se habla de que algo es interpretable cuando es en principio comprensible de manera *semántica*. Ya hemos usado este término anteriormente. Con más detalle: lo semántico hace referencia a una estructura de lenguaje proposicional. Aquello que no es proposicional, no tiene una estructura general con un set de conectivos específicos a ese lenguaje y con un conjunto de formas determinadas en que se formulan sus proposiciones. Si algo no es interpretable, es porque los datos que lo conforman no están estructurados con esa estructura general ni con conectivos específicos ni con un conjunto de formas determinadas de formulación. El conjunto de los pesos y sesgos de una red neuronal multicapa no es semántica justamente por estos motivos.

Otra forma de dividir aclarativamente los métodos de interpretación es (2) según qué tan generalizable sea su aplicación entre modelos: el método puede haber sido diseñado (2a) para aplicarse sobre *un* modelo en específico (*model-specific*), o (2b) puede funcionar con cualquier clase de modelo[25] (*model-agnostic*) (Molnar, 2022, p.22). Estos últimos son necesariamente *post hoc* y, también en consecuencia, necesariamente externos. Es decir, no pueden explorar elementos internos al modelo, sino solo sus *inputs* y *outputs*, ya que, de lo contrario, no podrían ser generalizables.

[25] Se les llama "agnósticos de modelo", ya que no dependen de un tipo específico de modelo para funcionar. Por simplicidad, aquí se les llamará también "modelos agnósticos".

También podrían dividirse según (3) la extensión a lo interno del modelo en que se aplican (Molnar, 2022, p.22). Es decir, si explican (3a) localmente o (3b) globalmente. Si la interpretación es local, significa que el método se concentra en una predicción individual, o en un aspecto específico de esa predicción. Si es global, entonces explica el modelo completo, tomando en cuenta todos sus parámetros y posibles predicciones (y no ya una sola en específico).

Finalmente, la forma más específica de dividir los métodos de interpretación, es según (4) el resultado obtenido una vez aplicado tal método (Molnar, 2022, p. 21). Este puede mostrar (4a) estadísticas resumidas de características —principalmente numéricas en naturaleza, en contraposición a las visuales—, (4b) estadísticas visuales de características —tales como los gráficos de dependencia parcial, que funcionan presentando de manera visual, no solo numérica, algún resultado de las estadísticas resumidas en (4a)—, (4c) elementos internos del modelo —como es el caso de los mapas de calor o los pesos aprendidos en modelos lineales—, (4d) datos individuales —que deben ser interpretables por sí mismos, como en el caso de predicciones de análisis de sentimiento textual, en las que se puede señalar una palabra (el dato individual) como la causante principal de la predicción— o, finalmente, el resultado puede mostrarse como (4e) un modelo inherentemente interpretable —o sea, se transforma alguna sección de un modelo *post hoc* en, por ejemplo, una regresión lineal que sea interpretable en sí misma, de modo que se puedan ver sus parámetros—. Esta división ayuda a distinguir la clase de modelos de interpretación que se utilizan comúnmente en la actualidad, aunque evidentemente podría (y debería) estar sujeta a cambios.

Las diferentes formas de clasificar modelos de interpretación ayudan a formar un panorama general de los métodos que se tienen hoy en día para hacer más transparentes los modelos de ML. A esto será oportuno sumar una enumeración de sus objetivos principales, de modo que se condense adecuadamente el fin del *sistema* que representa el área de IML. Los modelos de interpretación en ML tienen, al menos, tres funciones relevantes: (1) validación de modelos, (2) depuración de modelos —mencionado con anterioridad— y (3) descubrimiento de conocimiento (Du *et al*., 2019, p. 73).

Incluso siendo muy preciso con los datos de entrenamiento a la hora de hacer predicciones, es necesario validar un modelo de ML para asegurarse de que no está permeado por sesgos ilegítimos, poco éticos o simplemente equivocados. Esto, en el caso de modelos de ML epistémicamente opacos, solo es posible mediante algún método de interpretación. Por ejemplo, al

aplicar un mapa de calor[26] a un clasificador de imágenes que señala con alta precisión si en una foto se encuentra un husky o un lobo, se evidencia que uno de los detalles que más influyen en la predicción del modelo es el fondo de la imagen, ya que en la mayoría de fotografías de lobos el fondo estaba dominado por nieve y, por lo tanto, por el color blanco (Ribeiro *et al*., 2016). La precisión del modelo era alta, pero se basaba meramente en una correlación, y no en una relación directa con el objeto que se intenta clasificar. Al mismo tiempo, esa precisión era con respecto al conjunto limitado de datos con los que había sido entrenado el modelo y con los que se evaluó. El método de interpretación ayuda a validar (o invalidar) el modelo, que puede ser preciso pero inadecuado: la precisión siempre va a ser local en tanto que se mide según ese conjunto de datos específico con que se entrenó y en tanto que se evalúa, al menos inicialmente, también mediante un conjunto de datos específico.[27]

Lo anterior es importante por dos razones: en primer lugar, se evidencia que la efectividad y/o precisión de un modelo no necesariamente implica su adecuación —entendiendo adecuación como algo que no debe ser relativo únicamente al conjunto de datos de que se dispone para evaluarlo, sino a su aplicación en conjuntos de datos completamente nuevos para el modelo—, ya que esta es solo accidental y en ciertos escenarios importantes podría resultar en funcionamiento completamente defectuoso. Por ejemplo, en el caso de que se aplique el clasificador del ejemplo en imágenes de lobos en zonas de conservación donde no hay nieve, en cuyo caso sería completamente inexacto y, por lo tanto, inadecuado. Y, en segundo lugar, porque es un detalle atinente al problema específico de esta investigación, que tiene que ver con la relación epistemológica que existe entre los modelos de ML y la producción científica. Los resultados que pueda tener un modelo de ML, una simulación, o cualquier otra herramienta que pueda servir para algún fin científico, nunca bastan

[26] Método visual de interpretación que muestra las zonas de una imagen (o las palabras de un texto, en el caso de un procesador de lenguaje natural) que resultan determinantes para la decisión final (*output*) del modelo interpretado.

[27] Una aclaración que puede ayudar a dilucidar la aparente contradicción de algo "preciso pero inadecuado" es la distinción entre el concepto de precisión y el de exactitud. La última se refiere a la cercanía de un valor con respecto de aquello que intenta medir. La precisión se entiende más bien como el nivel de dispersión de los valores de medición. Por ejemplo, si en las imágenes del caso de los lobos se obtuviera en una ocasión el valor "cánido", la exactitud de esta predicción sería absoluta, ya que todas las imágenes muestran un lobo o un perro (husky), cánidos ambos. Sin embargo, si a la siguiente predicción se obtuviera el valor "lobo" y aún a la siguiente se obtuviera "husky" y después "animal", "mamífero", etc. podríamos decir que se trata de un modelo impreciso, ya que la dispersión de valores sería más alta que si solo se limitara a "husky" o "lobo", pero seguiría siendo exacto, en tanto que todas las respuestas son certeras respecto de aquello que intenta determinar. Es fácil imaginarse entonces cómo algo puede ser preciso pero inexacto, y por lo tanto inadecuado para algún fin en particular. Ya anteriormente se había hablado del concepto de precisión y sus diferentes acepciones, en caso de querer profundiza sobre el tema.

por su cuenta (a menos que se trate de un modelo inherentemente interpretable), sino que siempre van a requerir de un método de interpretación que los haga transparentes y que los valide. Esto evidencia claramente el problema del uso de modelos complejos de ML para fines científicos, ya que resultados sin una explicación no parecen calzar con el concepto de ciencia, incluso siendo muy precisos[28]. Sin embargo, al mismo tiempo permite vislumbrar posibles soluciones.

Continuando con los objetivos del IML, la depuración de los modelos de ML también juega un papel necesario. Se trata de comprender por qué un modelo está produciendo resultados que no son los deseados, lo cual se dificulta en mayor medida cuando es complejo y opaco. Para utilizar el mismo ejemplo del lobo, se podría buscar entender por qué el modelo se está enfocando en el fondo de las imágenes antes que en los detalles del animal en sí mismo que se espera clasificar. Una buena opción consistiría en aplicar ejemplos conflictivos (Papernot *et al*., 2017), que son una clase de modelo de interpretación *post hoc* que consiste en manipular al modelo para obtener el resultado deseado con imágenes que para un humano de ninguna forma podrían estar representando a un lobo en este caso. También lo contrario es posible: manipular la imagen para que, a pesar de que sea a todas luces (humanas) un lobo, el resultado sea cualquier otra cosa. Esto permite tener una idea más clara de qué es lo que se debe arreglar en el modelo original para producir, ahora sí, resultados menos manipulables y mucho más adecuados. Este es, no obstante, un objetivo que satisface necesidades más técnicas.

Por último, el otro objetivo significativo de IML es el del descubrimiento de conocimiento. Los modelos opacos de ML, como se ha analizado, pueden ser precisos, efectivos y hasta adecuados por sí mismos. Pero si se dejan por su cuenta, se está pecando de ineficiencia, pues se estaría ignorando una gran cantidad de valor que se puede extraer solo del hecho de comprenderlos mejor, de hacerlos más transparentes. Ya se vio que puede ayudar a convertir los resultados de un modelo en algo que pueda ser considerado científico al ser validados. Pero el beneficio va más allá: un modelo transparente puede ser *explorado* (Cichy & Kaiser, 2019, p. 9) para encontrar nuevos caminos para la producción de conocimiento. Uno opaco, por otra parte, no ofrece más que sus *outputs*. Pero al hacer más accesible el funcionamiento del modelo, ya sea a lo interno o de forma *post hoc*, se puede buscar perspectivas (*insights*) que no se preveían o que no se presupuestaban y

[28] La acepción aquí utilizada de precisión hace referencia al desempeño del modelo. Se basa en (Alpaydin, 2010, p. 492), donde se define como el número de registros recuperados y relevantes dividido entre el número total de registros recuperados.

que pueden hablar del funcionamiento de algún fenómeno. La investigación se extenderá más sobre este tema en secciones posteriores.

Además de estas tres finalidades, de naturaleza más epistémica, sería posible agregar una más de corte social y/o ético: la generación de confianza en los modelos de ML. Este es un objetivo de suma importancia en investigaciones recientes (Chatzimparmpas *et al*., 2020; Rescher & Kaminski, 2019; Rudin & Ustun, 2018; Schmidt & Biessmann, 2019; Yang *et al*., 2020; Yin *et al*., 2019), que pretenden establecer las condiciones necesarias para lograr esa confianza en los usuarios, tanto expertos como inexpertos, a la hora de utilizar modelos de ML para la toma de decisiones relativamente importantes. A pesar de que no se vaya a abordar este tema con mayor detalle, cabe mencionar que tiene un lugar de importancia en esta área de estudio.

2.3 LIME y Anclas: dos ejemplos de modelos de interpretación

Con la categorización realizada en la sección anterior, es posible presentar tres ejemplos de modelos de interpretación que muestran las características centrales de esta clase de herramientas, así como sus capacidades y sus limitaciones. Con ellos, se cumplirá la función doble de mostrar el estado de la cuestión con respecto a IML y, al mismo tiempo, señalar el punto de partida para comprender con mayor detalle y concreción la pregunta que se buscará responder en el tercer capítulo: si las herramientas de análisis de datos opacas eliminan la necesidad de teorización y la búsqueda de relaciones causales en la ciencia, o si, más bien, no pueden ser consideradas científicas del todo.

Para empezar, LIME (*Local interpretable model-agnostic explanations*) es un algoritmo que genera un modelo sustituto local. Es decir, su método de interpretación consiste en sustituir el modelo que interpreta por otro local, lineal, y, por lo tanto, cuya comprensión se vuelve asequible. Se trata de un modelo que aplica a cualquier otro (*model-agnostic*, o del tipo 2b, según la categorización de la sección anterior), lo cual significa que, por necesidad, es también post hoc y se centra en datos externos al modelo interpretado (sus *inputs* y sus *outputs*, no sus parámetros o estructura).

La interpretación se da en cinco pasos (Molnar, 2022;): (1) se selecciona la predicción específica por explicar, (2) se perturba el conjunto de datos —es decir, se modifican los datos utilizados como *inputs* cambiando las variables individualmente en alguna medida relevante— y se

utilizan como *inputs* nuevos para obtener *outputs* a partir de ellos, (3) se comparan con los *outputs* obtenidos con el conjunto original para dar mayor peso a los resultados cercanos a la predicción específica por explicar, (4) se entrena un modelo ahora de suyo interpretable (por lo general lineal) con base en el conjunto de datos con sus variaciones y los pesos incrementados respectivos, lo cual es realizable debido a la localización de las predicciones modificadas (no se toma el modelo globalmente) y (5) finalmente se obtiene la explicación a través del nuevo modelo de suyo interpretable.

De modo más conciso: se toma un modelo no-interpretable (probablemente por ser no-lineal y/o poseer varias capas escondidas) y se simplifica (o reduce) a uno interpretable (lineal, sin caja negra) utilizando la predicción específica que se quiere explicar como eje para la simplificación. La explicación consiste en mostrar las características que más influyeron en la predicción específica revisada, y qué tanto peso tuvieron sobre ella, de modo que se pueda inferir el funcionamiento del modelo opaco en ese punto local en particular. Entre las ventajas de esta reducción, está el hecho de que sea agnóstica con respecto a distintos modelos: al depender solo de los *inputs* y los *outputs*, y no de lo que sucede al interior de la caja negra, se puede aplicar a cualquier modelo, sea complejo o simple, sea de datos tabulares, de texto o de imágenes (Ribeiro *et al*., 2016, 1144). Acompañado a lo anterior, están las ventajas generales que se producen por poseer una explicación más o menos adecuada de la caja negra: mayor confianza del usuario, la posibilidad de optimizar el modelo original en caso de un funcionamiento ineficiente o ineficaz, etc.

Sin embargo, las desventajas también son múltiples y significativas. Una de las más influyentes es la que respecta a las diferentes fuentes de incertidumbre que agrega al proceso que va desde la predicción del modelo por interpretar, hasta su intento de comprensión por parte de una persona (Zhang *et al*., 2019). Si ya existe incertidumbre en la predicción de un modelo, LIME agrega aún más por (i) la aleatoriedad en el procedimiento de muestreo (explicado anteriormente en (2)), lo cual genera con bastante frecuencia diferentes interpretaciones al utilizar LIME en repetidas ocasiones sobre un mismo punto de datos. También agrega incertidumbre por (ii) la variación de la proximidad de las muestras, que produce diferentes interpretaciones dependiendo de la cercanía o lejanía de los datos muestreados entre sí (se trata de un parámetro modificable y que depende de si se busca mayor globalidad o localidad en la interpretación). Y, finalmente, por (iii) la variación en la credibilidad asignada al modelo original que producen las interpretaciones de

LIME: si se utiliza sobre diferentes predicciones de un modelo, en algunas habrá explicaciones que generen confianza, en muchas otras todo lo contrario (Zhang *et al*., 2019).

Para contrarrestar este influjo adicional de incertidumbre, se ha propuesto el concepto de "robustez" de las explicaciones como medida cuantificable del nivel de similitud de interpretaciones obtenidas sobre *inputs* semejantes (Alvarez-Melis & Jaakkola, 2018). Un modelo de IML debería tener como una de sus prioridades el llegar a un nivel aceptable de robustez —que, para *inputs* semejantes, las interpretaciones no sean demasiado dispares—. El problema con LIME y otros modelos agnósticos, basados por lo general en el proceso de perturbación de datos, es justamente que no presentan un nivel aceptable de robustez (Alvarez-Melis & Jaakkola, 2018): mientras que pequeñas perturbaciones no generan cambios en la predicción del modelo original, sí provocan modificaciones significativas en la explicación del modelo que interpreta.

Además, para socavar aún más la confianza producida por modelos agnósticos como LIME, se ha demostrado la posibilidad de manipular los resultados de un modelo de ML de caja negra mediante ejemplos conflictivos (Papernot *et al*., 2017), tal y como se explicó anteriormente (sección 2.3). Sin embargo, esto significa que también es posible realizar ataques conflictivos a los modelos que interpretan, no solo a los modelos que son interpretados (Bordt *et al*., 2022; Slack *et al*., 2020). En otras palabras, se puede modificar un modelo para que funcione con sus sesgos normalmente, pero que, a la hora de perturbar los datos para ser interpretado, funcione sin esos sesgos, de modo que se produzca una explicación inocua a pesar de que tenga sesgos potencialmente perjudiciales. Esto implica una incapacidad de coordinación entre la explicación y lo que explica, lo cual se torna más pernicioso en casos de contextos de conflicto, donde quien explica y quien recibe la explicación tienen finalidades opuestas. Más adelante se aclarará con mayor detalle estos conceptos y situaciones.

Finalmente, también se ha demostrado que, mientras que las explicaciones de LIME suelen generar más entendimiento sobre el modelo que en caso de no haber explicación del todo (aunque no en todos los casos), no generan en el usuario la capacidad de predecir con niveles consistentes de precisión lo que va a hacer el modelo una vez que ha sido explicado (Ribeiro *et al*., 2018). Esta es una forma común de probar la comprensión de los usuarios sobre el funcionamiento de un modelo: se muestra una explicación de su funcionamiento a un grupo de personas (con diferentes grados de familiarización con el área de ML) y sus resultados, y a un grupo de control no se le muestra explicación alguna, sino únicamente los resultados. Luego, se muestran *inputs* y se les pregunta a

las mismas personas si se sienten capaces de predecir el *output* que tendrá el modelo. De esta forma, se miden dos parámetros: cuántas veces intentan predecir, y qué tan precisa y correcta es la predicción. Con LIME, se hacen más predicciones que sin explicación del todo, pero, como se dijo anteriormente, su precisión y corrección es inconsistente (Ribeiro *et al*., 2018, 1533).

Existe otra herramienta de IML que tiene el resultado inverso con respecto a ese experimento: las anclas (*anchors*). Las diferencias en el modo de explicación que ofrecen tienen como producto una menor cantidad de intentos de predicción por parte de los usuarios, pero con una mayor precisión (predicen correctamente lo que el modelo interpretado va a tener como *output*). Podría parecer en primera instancia que la menor cantidad de intentos de predicción es algo negativo: una herramienta de interpretación ideal tendría una cobertura completa, en el sentido de que sus explicaciones producirían en los usuarios un entendimiento tal, que siempre se sentirían confiados a predecir el *output*. Sin embargo, se puede ver desde una luz positiva: mientras que la cobertura no es tan amplia como con LIME, el nivel de certidumbre en los casos en que se intenta sí es mayor —o, de otro modo, permiten mayor intensión, pero menor extensión—, lo que evita falsos positivos que podrían ser problemáticos por generar malinterpretaciones (una confianza engañosa sobre cómo funciona el modelo).

Las explicaciones de anclas son modelos agnósticos que funcionan con el mismo proceso de cinco pasos de LIME revisado anteriormente, con la diferencia principal en el quinto paso: el modelo interpretable busca reglas de 'si/entonces' que fijen los valores específicos de las características de una instancia que mantengan la predicción a pesar de que se cambien todos los demás valores no fijados. Es decir, en lugar de cuantificar el efecto específico que tuvieron todas (o muchas de) las características de la instancia en el resultado de la predicción, tal y como lo hace LIME, se "anclan" las características suficientes (y necesarias) para llegar a ese mismo resultado (Ribeiro *et al*., 2018, 1527). Por ejemplo, piénsese un algoritmo que predice la función gramatical de las palabras en una oración. Para explicar cómo predice el modelo la palabra "juego", podría decir: **si** la palabra previa es un pronombre, **entonces** "juego" es un verbo; **si** la palabra previa es un adjetivo, **entonces** "juego" es un sustantivo (se trata de un ejemplo similar al de Ribeiro *et al*. (2018, tabla 1)). Una explicación de LIME, por otra parte, mostraría cuánto influyó cada palabra de la oración, y no solo la previa, en la predicción final.

Las explicaciones de ancla resultan más intuitivas que las de LIME, en tanto que utiliza formulaciones condicionales que le dan una apariencia de causalidad. Como se verá más adelante,

la causalidad (no la correlación o las asociaciones estadísticas) es la forma de explicación más comprensible para el humano promedio (Miller, 2019, 6). Puede parecer extraño en primera instancia, ya que la explicación de LIME aporta, de suyo, mayor cantidad de información. Sin embargo, muchas veces es justamente ese detalle el que limita la comprensión precisa del funcionamiento de algún sistema. Ahora bien, por más intuitivas que resulten las anclas, tienen también desventajas y debilidades importantes.

En primer lugar, como se insinuó ya con anterioridad, el hecho de que los usuarios se atrevan en menos oportunidades a hacer predicciones sobre el modelo interpretado podría ser percibido como positivo desde la perspectiva de minimización de falsos positivos. Los usuarios muestran una comprensión más intensional del modelo y prefieren responder "no sé" cuando la extensión de la explicación no cubre el caso (Ribeiro *et al*., 2018, 1533), antes que hacer el intento de responder. Sin embargo, la falta de extensión de esta clase de explicación es, en sí misma, un problema, especialmente cuando la predicción está en los límites de la capacidad del modelo. En estos casos, el ancla resulta ser excesivamente específica, ya que los factores necesarios y suficientes para la decisión terminan siendo muy numerosos, de modo que deja de gozar de los beneficios que se comentaron anteriormente.

Y, en segundo lugar, los problemas que tenía LIME por requerir de un proceso de perturbación de datos, tales como el aumento innecesario de la incertidumbre y un potencial incremento del ruido, así como la posibilidad de alterar intencionalmente las explicaciones según sea necesario, también afectan a la explicación de anclas, ya que se basan en ese mismo proceso. Este último no es un problema menor, tanto para el uso de modelos complejos de ML (como redes neuronales de múltiples capas o bosques aleatorios) en la toma de decisiones que afecte social, política o económicamente a los individuos, como para su uso en la ciencia, que es el tema de esta investigación.

Un contexto de cooperación es uno en el que todos los involucrados tienen intereses paralelos. Su contrario, un contexto de conflicto, es uno en el que los involucrados tienen intereses opuestos (Bordt *et al*., 2022). En instancias del primer tipo de contexto, la explicación que provean los modelos de interpretación *post hoc* tiene una clara utilidad y no genera desconfianza externa. La desconfianza provendría de qué tan precisos son esos modelos en sí mismos, no de si se están utilizando con alguna intención ulterior en mente, más allá de la compartida por todos los involucrados. Como ejemplo, se puede considerar el caso de una analista de datos que aplique un

modelo de interpretación sobre una caja negra con el fin de depurarla: ella es quien aplica el modelo para obtener una explicación y, al mismo tiempo, es quien busca o recibe la misma. Los objetivos de quien aplica el modelo de IML y de quien recibe su resultado se alinean: "ambos" quieren depurarlo. En estos casos, los modelos *post hoc* suelen ser muy útiles y beneficiosos (Bordt *et al*., 2022).

No ocurre lo mismo con los contextos de conflicto, en los que los involucrados —quienes dan la explicación y quienes la reciben— tienen fines opuestos. En estos casos, los modelos de interpretación pueden ser fácilmente manipulados para obtener el resultado esperado por quien los aplica. Varias de sus características naturales son responsables de que esto sea el caso. Por ejemplo, no existe tal cosa como una explicación *real*, *subyacente* y única de por qué un *input* generó un *output* específico, ya que siempre hay un conjunto indeterminado de razones, todas con pesos distintos. Lo único que hay son las meras aproximaciones a esas razones y pesos que los modelos de interpretación como LIME, anclas y SHAP ofrecen (Bordt *et al*., 2022). Esto será explorado con mayor detalle en la siguiente sección, en la que se considerará esta característica como fundamental para comprender el rol de los modelos complejos de ML en la ciencia. Y, si se combina este hecho sobre la inexistencia de una explicación subyacente, con el otro hecho de que los diferentes modelos de interpretación producen distintas explicaciones (como se vio, por ejemplo, en el caso de LIME y anclas), lo que se obtiene es la posibilidad de quien da la explicación *escoja* qué explicación da según sea su conveniencia[29]. Después de todo, no existe una explicación fundamental que sirva de contraste con el resultado.

Bordt *et al*. (2022) consideran conflictivos casos como el de la decisión de un banco de otorgar o no una deuda, la decisión de una universidad de si admitir o no a un estudiante y la decisión sobre si una persona debe permanecer o no en la cárcel por peligro de reincidencia. Sin embargo, coloca el caso del descubrimiento científico en el lado de la cooperación. Esta postura depende del modo en el que se esté considerando la ciencia: desde una perspectiva teórico-ideal, o desde una perspectiva sociológica. Evidentemente, los autores están considerando la primera opción. La segunda opción, no obstante, podría resultar en la visión del descubrimiento científico como un ejemplo de contexto conflictivo. Como ha quedado claro en la última década tras la crisis

[29] Esta elección conlleva, no solo el modelo de interpretación utilizado para explicar la predicción, sino los parámetros según los cuales se ejecuta ese modelo de interpretación, tal y como se explica con mayor detalle en (Bordt et al., 2022, 9).

de replicación, los diferentes incentivos y desincentivos[30] que tienen las personas que producen ciencia, han causado un ambiente en el que no es poco común que se manipule de una u otra forma una investigación para que genere los resultados deseados, no necesariamente los verdaderos (Cockburn *et al*., 2020; Maxwell *et al*., 2015; Shrout & Rodgers, 2018; Wiggins & Christopherson, 2019). El uso de un modelo de caja negra cuya explicación es mediada por un modelo de interpretación podría prestarse para estos fines. Después de todo, quienes dan la explicación y quienes la reciben tienen fines distintos: unos quieren generar y validar (aunque sea de forma espuria) un resultado particular que beneficie su investigación en algún sentido, otros quieren simplemente saber cómo se llegó al resultado obtenido.

A pesar de lo anterior, el foco de esta investigación no son los posibles efectos en términos sociológicos que tengan los modelos opacos sobre la producción de conocimiento científico. Se trata más bien de la ciencia en el sentido teórico-ideal mencionado más arriba, o, si se quiere, desde la perspectiva meramente teórico-epistemológica. La siguiente sección, por lo tanto, lidia con el asunto de los límites de los modelos de interpretación como solución a la opacidad de los modelos complejos, especialmente considerando el campo de la producción científica.

2.4 Incompletitud y coordinación: problemas fundamentales de la interpretabilidad de modelos

Hasta el momento se ha descrito el problema de la opacidad epistémica de los modelos complejos de ML. Aunado a esto, se presentó IML, área de la ciencia de datos que se dedica a generar modelos especializados en brindar explicaciones sobre el funcionamiento de esos modelos opacos. La finalidad de esta última consiste en mitigar esa opacidad epistémica, provocar mayor transparencia y, consecuentemente, hacerlos más comprensibles. Esto tiene muchísimas ventajas de diversos tipos: mayor confianza en modelos de ML que toman decisiones importantes sobre individuos y sociedades, mayor capacidad de depuración para quienes trabajan en ellos y, para los científicos, mayor entendimiento de los resultados obtenidos para la generación de hipótesis y teorías con base en sus resultados. Sin embargo, estos modelos de IML, de los cuales se dieron dos ejemplos más arriba, tienen varios problemas epistemológicos que ponen en duda si logran su

[30] Salariales, laborales, de estatus, etc.

cometido de abrir la caja negra, o no. A continuación, se plantearán algunos de esos problemas epistemológicos y se pondrán en relación con la práctica científica.

Primero, hay que recordar los objetivos que tiene IML como campo: validar, depurar y descubrir (Du *et al*., 2019)[31]. ¿Cumplen los actuales modelos de IML con estos objetivos? ¿En qué medida? Y, ¿deberían ser estos los objetivos y no otros? Por lo revisado en la sección anterior, se podría responder a la primera pregunta diciendo que cumplen hasta cierto punto con cada objetivo, pero que la manera en que lo hacen no es lo suficientemente consistente para tomarlos como acabados, lo cual supone un problema de confianza ulterior.

Piénsese, por ejemplo, en los problemas que se discutieron en la sección anterior al implementar un modelo como LIME: el incremento de incertidumbre sobre la ya existente en el modelo por interpretar es demasiado elevado. O, en otras palabras, la robustez de los modelos de IML que funcionan perturbando los datos no es lo suficientemente alta para que sean realmente confiables (Zhang *et al*., 2019; Alvarez-Melis & Jaakkola, 2018). ¿Cómo es posible validar consistente y precisamente un modelo si este es el caso? Y se está hablando solo de un incremento de incertidumbre, el cual es en principio cuantificable. Es decir, al menos se tiene una idea de la magnitud de imprecisión[32] del modelo de IML. Hay, no obstante, otros problemas que no son siquiera cuantificables y que también ponen en duda la confiabilidad de los modelos de interpretación.

Para comprender lo anterior, cabe hacer uso de una distinción conceptual que resulta muy efectiva en este campo de conocimiento: la incertidumbre, por un lado, y la incompletitud, por el otro (Doshi-Velez & Kim, 2018). La incertidumbre —en su sentido estadístico— se puede entender como varianza cuantificada. Es decir, se trata de algo sobre lo que se puede razonar y que puede ser en principio calculado. Los problemas que se acaban de señalar sobre la perturbación de datos son de esta clase. La incompletitud, por otro lado, hace referencia a un sesgo no cuantificable (o no cuantificado). Doshi-Velez y Kim (2018) utilizan esta distinción para justificar la necesidad de los métodos de explicación *post hoc* sobre modelos de ML: al padecer muchos modelos de una incompletitud en la formulación de los problemas que buscan resolver, las explicaciones producidas

[31] Watson y Floridi (2020) hablan de auditar, validar y descubrir, donde auditar toma el papel ético que tiene "validar" en (Du et al., 2019), y validar toma el papel más técnico de "depurar". Sin embargo, las descripciones son prácticamente las mismas.

[32] Cabe aclarar que los rangos aceptables o inaceptables de imprecisión o de incertidumbre suelen ser determinados de manera relativamente arbitraria y no por alguna razón subyacente.

por IML podrían ayudar, no a resolver, pero sí a hacer visibles esas brechas de incompletitud para que no pasen desapercibidas. Posteriormente podrían intentar ser resueltas. Sin embargo, se podría argumentar que la incompletitud no se queda únicamente en los modelos de ML por interpretar, sino que varios de los mismos modelos de IML cuya función es reducir esa incompletitud, más bien la poseen por su cuenta y la están incrementando, generando un problema de confianza en la comunicación de lo que desean explicar.

Para exponer con mayor detalle este argumento se necesitan tres consideraciones: (1) exponer la incompletitud en los modelos de ML (no interpretados), (2) exponer la incompletitud en los modelos de IML y (3) exponer el problema lógico de *los dos generales* como herramienta para entender la imposibilidad comunicativa de confianza entre los modelos de ML y los de IML.

Sobre (1) la incompletitud en ML: muchos modelos, especialmente los que no son de suyo interpretables y tienen funciones complejas, no puede abarcar la totalidad de casos en que pueden fallar (Doshi-Velez & Kim, 2018). Piénsese, por ejemplo, en el caso de algoritmos para la conducción automática de vehículos de transporte. Estos pueden ser muy precisos, pero siempre en un conjunto constreñido de casos similares a aquellos con los que fueron entrenados. Pero son tantos los eventos posibles que pueden ocurrir cuando se conduce —el sistema en el que se realiza la conducción tiene una cantidad, no infinita, pero sí incuantificable de elementos que pueden influir en una infinidad de maneras—, que no es factible considerarlos todos y cada uno. Se puede aplicar reglas de generalización, pero estas no van a cubrir la totalidad de desenlaces posibles. Como se dijo, esos desenlaces y los elementos que los causan no son cuantificables en principio, por lo que no se puede hablar de ellos como productores de 'incertidumbre' en el sentido estadístico. Lo que producen es, más bien, incompletitud: el sistema que se intenta modelar es tan complejo, que es imposible saber qué tanto le falta a este para considerarse 'completo'.

El anterior es solo un ejemplo, pero muchos tipos de problema relacionados con modelos de ML tienen su punto de origen en este concepto. Las consideraciones sobre seguridad, ética y los objetivos mismos del modelo se ven atravesados por él (Doshi-Velez & Kim, 2018), ya que, al ser asuntos complejos, se introducen errores y sesgos no previstos y no cuantificables que tampoco permiten garantizar un uso seguro, moral y adecuado de ellos. Sin embargo, aquí interesa un tema en específico, y es el de la comprensión científica.

Sobre este tema se profundizará con más detalle en el siguiente capítulo, pero sirva aquí como una pequeña introducción y para contextualizar las deficiencias de IML. Una de las principales

funciones de los modelos complejos de ML en la ciencia, es la de la exploración y el descubrimiento de patrones relevantes (Cichy & Kaiser, 2019; Du *et al*., 2019; Watson & Floridi, 2020). Es decir, los modelos de ML no son los que exploran, sino que simplemente presentan un conjunto de patrones difíciles o imposibles de descubrir para un ser humano. La tarea de este último es explorar esos patrones y escoger los que sean relevantes —la relevancia es un aspecto en el que el ser humano es adepto, mientras que las computadoras (todavía) no lo son—. Sin embargo, el conocimiento científico responde a un sistema muy complejo, de modo que la incompletitud es casi parte fundamental de este. Por ejemplo, cuando se utilizaba el modelo aristotélico de la física, no había forma de saber que iba a haber otro modelo más exitoso en el futuro: el sesgo y la magnitud de los errores que producía eran incuantificables. Actualmente tampoco es posible saber si los modelos actuales serán reemplazados por otros mejores, pero al menos se tiene una mayor claridad sobre lo que constituye una aproximación y los errores que puede generar. En consecuencia, cuando se utilizan modelos complejos de ML para hacer ciencia, se está contribuyendo con la incompletitud de la comprensión científica. ¿Se trata de un intercambio beneficioso para el avance científico? ¿Hay alguna forma de que no sea un intercambio del todo?

(2) La incompletitud en IML es una potencial respuesta a las preguntas anteriores. De hecho, como se dijo anteriormente, los mismos autores que proponen la distinción entre los conceptos de 'incompletitud' e 'incertidumbre', lo hacen con la finalidad de señalar que IML es una solución posible a los problemas de incompletitud producidos por utilizar modelos complejos de ML (Doshi-Velez & Kim, 2018). Aquí, no obstante, se argumenta que muchos de los modelos de IML que buscan solucionar ese problema, sufren del mismo, y que esto los hace, en principio, incapaces de solucionar el problema inicial. Antes de delinear el argumento, sin embargo, es imperativo hacer una precisión: el problema de la incompletitud aplicado a IML, igual que cuando es aplicado a ML, no es algo de lo que todos los modelos participen. Cuanto más simples sean, tantos menos factores pueden influir sobre los resultados. Justamente es esto lo que los hace simples. Lo mismo se puede decir sobre los modelos de IML. No obstante, los modelos de explicación *post hoc*, los cuales son en su mayoría agnósticos (no solo aplican para un modelo específico), sí son víctima de este problema a pesar de que sean simples y simplificantes. A continuación, se aclara por qué.

Tal y como admiten Doshi-Velez y Kim (2018), la incompletitud de los problemas puede provenir de varias fuentes: una incompletitud de los potenciales *inputs* que alimentan al modelo, la incompletitud de los potenciales *outputs*, la incompletitud de problemas éticos que se pueden

desprender del uso del modelo, la incompletitud en la comprensión del funcionamiento del modelo mismo, la incompletitud de los dominios especificados del modelo, etc. (p. 12). Se puede aceptar también que los modelos de IML que son agnósticos —es decir, que funcionan solo con los *inputs* y los respectivos *outputs* de un modelo, pero no con su armazón interna—, al utilizar el método de perturbación de datos, no logran alcanzar un nivel aceptable de robustez (Alvarez-Melis & Jaakkola, 2018). Esto se debe a que la aleatoriedad de la elección de muestras de la que depende ese proceso, sumada a la aleatoriedad en el nivel proximidad de los *inputs* que se escogen para generar el modelo lineal local que permite la explicación, generan una varianza muy alta (Zhang *et al*., 2019, pp. 2-3). Esta varianza al ser cuantificada es incertidumbre, no incompletitud. Y desde cierta perspectiva lo es, ya que se puede calcular y cuantificar. Pero sí pasa a ser un problema incompletamente especificado cuando se piensa desde la explicación resultante: mientras que se puede cuantificar la varianza en factores específicos del proceso de muestreo, no se puede cuantificar cuánto afecta esa varianza en la adecuación de la explicación, a pesar de que se sabe que sí la afecta.

Para ejemplificar el argumento anterior, se ha señalado que la cuantificación de incertidumbre en modelos de DL es problemática debido a cuatro razones principales: (1) la ausencia de una teoría en dominios específicos que permita fundamentar esas incertidumbres (este es el caso, por ejemplo, en decisiones asistidas por ML en medicina), (2) debido a la ausencia de modelos causales (ya que DL se centra en meras correlaciones, las cuales pueden definitivamente brindar predicciones adecuadas, pero limitan las conclusiones que se puedan sacar de esas predicciones), (3) debido al altísimo costo computacional de aplicar los modelos de DL, lo cual implica un costo aún mayor y poco razonable para calcular las incertidumbres en esos modelos, lo cual requeriría correrlos en múltiples ocasiones, y (4) debido a lo sensibles que son esos modelos a pequeñas imperfecciones en los datos (Begoli *et al*., 2019). Este último motivo es el que más interesa aquí: los enfoques basados en datos, los cuales usualmente se relacionan con *Big Data* y con DL, utilizan “sutiles correlaciones multivariantes para mejorar sus predicciones” (Begoli *et al*., 2019, p. 4). El problema para modelos de DL, es que los datos de entrenamiento suelen ser imperfectos (con errores y elementos faltantes), lo cual termina confundiendo las predicciones y generando resultados que no logran cumplir con su finalidad y el nivel de precisión requerido. El problema para modelos de IML que intentan explicar a estos últimos, es que sus datos de entrenamiento son perturbaciones de esos mismos datos que ya daban problemas, las cuales incrementan aún más, y en una cantidad no cuantificable, la incertidumbre (lo cual la transforma en incompletitud).

Si se acepta la conclusión de este argumento, se acepta que, en su estado actual, los modelos de IML no son suficientes para superar el problema de incompletitud de los modelos complejos de ML. Al mismo tiempo, esto significa que sus objetivos no se cumplen a cabalidad. En el tercer capítulo se profundizará más específicamente sobre el objetivo relacionado con el descubrimiento científico, y se analizará con más detalle lo que implica que no se logre cumplir a cabalidad para la aplicación de ML en la práctica científica. Sin embargo, cabe recordar que este argumento se refiere específicamente a los modelos *complejos* de ML y los de IML que buscan dar la capacidad de interpretarlos a través de explicaciones. Como se dijo en la sección 2.3, hay autoras que plantean como deber moral el hacer uso de modelos simples, interpretables en sí mismos, de ML en contextos en que esté en juego la vida o el sustento de una persona (Rudin, 2019). Pero el deber podría ser considerado, no solo moral, sino también epistémico —asumiendo que se considere al menos un área de lo epistémico como normativa—: si los modelos complejos de ML producen problemas de incompletitud como los señalados arriba, y los modelos de IML que intentan solucionarlo también participan de esos problemas —independientemente de si todavía no son suficientes para solucionarlos, pero podrían llegar a serlo, o si en principio no son capaces de llegar a serlo del todo—, entonces no solo se ven afectados los modelos que tienen alguna influencia sobre la vida de las personas, sino también los que se utilizan para la producción del conocimiento, ya que no se podría estar seguro de que se está produciendo del todo. Los modelos clásicos de estadística no se verían enfrentados a este problema, ya que no habría falta de coordinación entre su interpretación (mediante explicación) y el modelo mismo: los modelos clásicos tienden a ser interpretables por sí mismos. Hay incertidumbre (varianza cuantificada), pero no incompletitud. El problema surge cuando se pierde el nivel semántico por la complejidad del modelo.

El argumento se ha llevado incluso más allá (Semenova *et al*., 2022), y se ha propuesto que, en la gran mayoría de los casos, es posible encontrar un modelo simple que cumpla adecuadamente con las funciones requeridas, por lo que es aún más imperioso buscarlo y aplicarlo en lugar de uno complejo. Para llegar a esta conclusión, se ha propuesto la idea de los *conjuntos de Rashomon* (Breiman, 2001): es el conjunto de modelos que obtienen funciones significativamente similares y que obtienen resultados razonablemente adecuados, a pesar de que utilicen algoritmos o hiperparámetros completamente distintos. Esta idea sugiere que, en un conjunto de Rashomon lo suficientemente amplio, siempre va a haber al menos un modelo de ML simple y, por lo tanto, por sí mismo interpretable. Si este es el caso, siempre debería ser escogido este último por sobre los demás, tanto por beneficios morales como epistémicos y prácticos. De este modo no se estaría

sacrificando precisión por interpretabilidad, ya que por definición son tan precisos como otros modelos del conjunto más complejos. Siendo así, tampoco habría necesidad de la explicación de modelos (y del campo de IML en general), siempre que se pueda encontrar ese modelo simple en el conjunto.

Ahora bien, la suposición de que siempre se puede encontrar un modelo simple e interpretable en el conjunto de Rashomon ha sido puesta en duda por la falta de fundamentos estadísticos o lógicos que permitan concluirla (Watson & Floridi, 2020). Lo más posible es que haya muchos casos en los que necesariamente se tenga que dar ese sacrificio de interpretabilidad para garantizar una mayor precisión o viceversa. Esto es especialmente cierto en un campo como el científico, donde muchas veces son los modelos complejos los que permiten un tipo nuevo de exploración que los modelos simples no permiten. Entonces, a pesar de que en muchos casos sea posible reemplazar modelos complejos por modelos simples sin una pérdida significativa de precisión, no se trata de una regla general. Por lo tanto, los esfuerzos de IML siguen siendo importantes y requeridos.

Pero en el estado actual de IML sigue siendo una preocupación el hecho de que no permite un nivel razonable de confianza, especialmente para el caso del objetivo de descubrimiento científico. Para comprender esta falta de confianza por la comunicación entre los modelos de ML complejos y los de IML, piénsese en el problema del ataque coordinado (Floridi, 2011) (también conocido como el problema de la falta bizantina o el problema de los generales bizantinos): hay dos generales con sus respectivos ejércitos, cada uno a un lado opuesto de un castillo que se encuentra en el punto medio entre ambos. Ambos tienen la misma finalidad: conquistar el castillo. Para ello, deben coordinar su ataque, ya que, si ataca un ejército sin la ayuda del otro, la defensa del castillo será inevitablemente superior y morirán todos en el intento, dejando al otro ejército por su cuenta, el cual también será, por lo tanto, insuficiente para lograr la meta. Ambos deben ponerse de acuerdo sobre el momento justo de ataque, para lo que disponen de un único medio de comunicación: mensajeros que van a caballo. Estos mensajeros, no obstante, solo pueden llegar al campamento del otro ejército pasando al lado del castillo, el cual vigila su exterior constantemente, ya que se encuentran en una zona montañosa, y durarían demasiado si intentaran rodear el paso directo. Así, cada vez que un mensajero se dirige al orto ejército, lo hace con una probabilidad significativa de morir en el intento puesto que los vigilantes del castillo intentarán detenerlo a toda costa.

El problema con esta situación es que estructuralmente *imposibilita* la coordinación del ataque. Cuando un mensajero del ejército A sale a informar al general B que la hora de ataque es tan pronto como salga el sol en el horizonte, el general A no tiene forma de saber si su mensajero llegó para dar la noticia. El general B, por lo tanto, se ve obligado a enviar otro mensajero en su nombre que *confirme* que llegó la noticia de la hora convenida del ataque. Pero el general B no puede estar nunca seguro de que su mensajero va a llegar al campamento del ejército A, por lo que necesitará, a su vez, de una nueva confirmación sobre la obtención de la confirmación inicial, y así *ad infinitum*. Esta imposibilidad comunicativa incluso se ha demostrado lógicamente (Fagin *et al*., 2003), y ha sido utilizada ilustrativa y analógicamente en diferentes campos para mostrar problemas de comunicación entre partes o conceptos[33].

Un caso como este, de información imperfecta y siempre incompleta, calza con el de la comunicación entre IML y ML: no se pueden coordinar de manera fidedigna las predicciones del modelo A con el modelo de explicación B, ya que siempre habrá una posibilidad de incompletitud en A y en B independientemente. Incompletitud que, por definición, no es cuantificable. ¿Cómo saber que el modelo B explica efectivamente el modelo A si, en principio, al presentar problemas múltiples de incompletitud, no se puede conocer la cantidad de ruido en la comunicación entre B y A?

Se podría objetar este argumento diciendo que IML no busca ajustarse de manera absoluta al modelo que explica: siempre que se intenta simplificar algo complejo, se pierde cierta cantidad de información. Lo que interesa es incrementar la comprensión del modelo complejo, no entenderlo de manera completa. Esto es cierto, una explicación nunca es una presentación tal cual del fenómeno que explica. Si fuera el caso, sería indistinguible una explicación de una mera descripción. Pero esta objeción no apunta al problema fundamental que señala el argumento anterior. No es relevante si la explicación que permite IML conlleva en sí misma una pérdida de información. Se acepta, de hecho, que esta pérdida de información es inherente a la explicación. Los elementos del conjunto de la explicación no calzan uno a uno con los elementos del conjunto del fenómeno explicado, por ponerlo en otras palabras, ya que siempre hay más elementos en el segundo conjunto

[33] El más común es en sistemas de computación distribuida, en los que los errores de las partes deben solucionarse de manera coordinada, pero no se puede tener información confiable sobre si cayeron en un error o no (Lamport et al., 1982). Floridi (2011), por su parte, utiliza el problema del ataque coordinado como método para mostrar que los conceptos de justificación y verdad son completamente independientes el uno del otro, de modo que la definición tripartita del conocimiento no puede esquivar de ninguna forma los contraejemplos de Gettier.

que en el primero: no es una función biyectiva. El problema radica, más bien, en el hecho de que no es posible saber si la explicación B es adecuada con respecto al modelo A. O sea, no afecta el hecho de que no se trate de una función biyectiva (entre la explicación y lo explicado), sino el hecho de que no se sabe si se trata de una función del todo, ya que no es claro de qué elementos están constituidos los conjuntos (problema de incompletitud).

Entonces, ¿permite IML el descubrimiento en la ciencia, o más bien lo tergiversa? Es posible pensar que, al paso al que avanza el campo, se llegue a superar muchos de los problemas aquí expuestos relativamente pronto. Sin embargo, para eso parece necesario tener en cuenta cuáles son los problemas que causa en primer lugar, cosa que no es común dentro de la literatura sobre el tema. Se debe evitar un avance a ciegas y sin considerar críticamente la manera en la que se está realizando. Al consistir en un análisis conceptual riguroso, la filosofía puede ayudar a esclarecer en alguna medida el camino. Por el momento, sin embargo, parece que IML no es la respuesta a los problemas del uso de ML y grandes conjuntos de datos en la producción de conocimiento científico.

En este segundo capítulo, se analizó el problema de la opacidad epistémica, conocido también como el problema de la caja negra, en ML. Se analizó el concepto mismo de aquello que constituye la opacidad epistémica y las diferentes razones por las que surge (tanto en modelos complejos de ML como en simulaciones computacionales complejas). Se revisó una solución potencial al problema en el campo de IML, para lo que se mostraron sus objetivos y un par de ejemplos de modelos agnósticos que buscan localizar las razones para el *output* final de un modelo original como forma posible de explicación de este último. Posteriormente, se vieron los problemas particulares de esos ejemplos, pero también los problemas generales de IML en general como método de explicación que permita construir conocimiento científico. En el siguiente capítulo se propondrán diversas definiciones de aquello que constituye el conocimiento científico, la explicación científica y la interpretación en general. Con ello se busca tener las herramientas conceptuales suficientes para poder determinar, finalmente, si el uso de modelos complejos de ML y de *Big Data* para su entrenamiento realmente lleva a conocimiento científico, o si simplemente dan respuestas sin una explicación, por lo que no podrían ser llamados esos resultados conocimiento científico (o conocimiento del todo).

Capítulo 3: ¿El fin de la teoría? La investigación basada en datos y los límites de la ciencia

Entre las múltiples utilidades que tienen los modelos complejos de ML, así como los grandes conjuntos de datos en formato computable que se poseen para alimentarlos, está la de la producción de conocimiento científico. Esto último es comprendido de manera amplia, ya que ML colabora en esa tarea de múltiples formas: no produce en sí misma el conocimiento científico, pero otorga a los investigadores un espacio nuevo para hacerlo. Tales herramientas, no obstante, son de una naturaleza tal, que generan cierto nivel de opacidad epistémica, como se demostró en el capítulo anterior, el cual podría entenderse como de un tipo novedoso, distinto de la opacidad epistémica social y técnica que ya ocurría en la ciencia antes de su aparición.

Como resultado, han surgido muchos cuestionamientos sobre la legitimidad científica de la aplicación de esas herramientas una vez que incluyen esa opacidad epistémica novedosa: ¿qué consecuencias se desprenden de su uso? ¿Significa esa opacidad epistémica una pérdida en la capacidad explicativa de la ciencia? Y si la respuesta es afirmativa, ¿no es una parte fundamental de la ciencia el hecho de ser explicativa? ML y su uso para el análisis de BD, además, han producido un tipo de método científico también novedoso (tal vez no en tipo, pero sí en intensidad): el acercamiento basado en datos (*data-driven science*). Se trata de una vía de producción de conocimiento científico que no requiere de teorización previa para llegar a resultados concretos, sino que utiliza el reciente y creciente acceso a datos computables para, a través de ML y/o simulación computacional, generar modelos que ofrezcan esos resultados de manera "automática". La razón por la que se obtienen esos resultados, sin embargo, puede, o no, ser opaca. ¿Se pueden considerar como conocimiento (científico) estos modelos o las conclusiones obtenidas mediante ellos cuando la razón es opaca?

Todas las preguntas anteriores requieren de definiciones claras de los elementos que utilizan. Será importante determinarlas antes de intentar dar respuesta. Se deberán sentar los conceptos de conocimiento y conocimiento científico, explicación y explicación científica y la ciencia como práctica. A partir de ellos, se podrá revisar con una estructura epistemológica explícita los problemas que surgen de la aplicación de ML para el análisis de BD específicamente en la práctica científica. Los problemas que se analizarán en las secciones subsiguientes se encuentran en un espectro que va desde problemas fundamentales y epistemológicos (opacidad epistémica), hasta problemas netamente prácticos, pero absolutamente relevantes, como el de la crisis de la

replicabilidad en la ciencia basada en ML. Todos ellos ponen en cuestión el hacer uso indiscriminado y poco cuidadoso y riguroso de estos métodos que se ha dado en años recientes, así como si eso que producen es, del todo, ciencia. Cabe aclarar, sin embargo, que no se busca con el presente capítulo negar activamente la *posibilidad* de hacer ciencia con ML a partir de BD. Se busca mostrar una serie de potenciales criterios para determinar cuándo una aplicación de ellas puede ser considerada adecuada a la ciencia y cuándo no.

Para trabajar los problemas señalados de manera detallada, este capítulo estará dividido en tres secciones. En primer lugar (3.1) se explorará el tema del conocimiento científico en general con el propósito de definir algunos conceptos centrales que se operacionalizarán en las secciones siguientes. Es decir, se definirá una red conceptual epistemológica que otorgará los criterios para el análisis posterior. Luego (3.2) se hablará sobre el rol de ML y BD en la práctica científica actual, para lo que se utilizarán ejemplos de uso en varias disciplinas. Finalmente, (3.3) se utilizará la estructura comprendida entre la red conceptual de (3.1) y los ejemplos de (3.2) para determinar si la aplicación de los métodos basados en datos mediante modelos complejos de ML significa una forma novedosa de comprender, no solo ya la opacidad epistémica, sino la práctica científica en general, así como sus consecuencias e implicaciones desde una perspectiva epistemológico-pragmática.

3.1 El conocimiento, la explicación y la práctica de la ciencia

La definición del concepto de conocimiento, que permanece en constante debate y actualización, es una de las tareas centrales de la epistemología. Por este motivo, es imposible dar una definición que esté libre a cabalidad de escepticismo y potenciales problematizaciones. De hecho, ni siquiera hay acuerdo sobre si es posible del todo definir el concepto o no. Sin embargo, se puede justificar el uso heurístico de una definición en particular siempre y cuando sea razonable y funcione dentro de un modelo específico del sistema. Para ello, en primer lugar y de forma sumaria, es necesario exponer el método de los niveles de abstracción (Floridi, 2008).

El método de los niveles de abstracción permite comprender las perspectivas específicas desde las que se construye una teoría. Aquello sobre lo que se hacen afirmaciones es un sistema. Para hacer afirmaciones sobre ese sistema, se construye un modelo[34], que es una versión

[34] El uso de la palabra modelo en esta sección es distinto del que se ha hecho en las secciones anteriores. Cuando hablamos de modelos en ML, hablamos de algoritmos entrenados, con parámetros ya determinados

simplificada de aquel, la cual considera únicamente los elementos que son relevantes para la finalidad con la que se esté utilizando. Esa finalidad es justamente lo que hace necesario definir los niveles de abstracción en los que se posiciona quien hace la afirmación. De esto no se desprende necesariamente que exista tal cosa como una jerarquía de los niveles de abstracción, sino que simplemente hay diferentes finalidades que interactúan de maneras diferentes con el modelo (Floridi, 2008). Por ejemplo, si se quiere investigar el problema de la opacidad epistémica en ML (el sistema) para proponer soluciones específicamente técnicas, se puede hacer un modelo que deje de lado cuestiones éticas, políticas o epistemológicas, y que solo incluya el ámbito técnico. Ese es el nivel de abstracción en que se ubica una investigación de tal naturaleza. Pero las afirmaciones que yo haga desde ese nivel de abstracción, no necesariamente funcionarán en un nivel de abstracción distinto: una solución técnica al problema de la opacidad epistémica podría no seguir ciertos estándares éticos necesarios para que sea satisfactoria. Comprendido así, el método de los niveles de abstracción tiene como consecuencia la implausibilidad de una afirmación que se haga sin una finalidad y, por lo tanto, sin un modelo específico sobre el cual se basa. Cabe aclarar, además, que no es consecuencia de este método un relativismo general: los modelos se construyen sobre la base de un sistema, y deben responder a este. Al mismo tiempo, aquello que se concluya de un modelo, debe ir acorde a la finalidad específica por la que existe: la finalidad y el sistema (el éxito interactivo entre modelo y sistema) son los factores que constriñen este método y no le permiten llegar a un relativismo ingenuo o absoluto.

Con el método de los niveles de abstracción en mente, ahora sí se puede justificar una definición de conocimiento como modo heurístico de facilitar el análisis que aquí se plantea. Como la finalidad de esta investigación es determinar cómo afecta el uso de modelos complejos de ML y BD a la producción de conocimiento científico, la definición de 'conocimiento' que se ofrece a continuación se relaciona con el conocimiento científico en particular y no considera características que podrían formar parte del concepto en contextos menos formales y más cotidianos. Por ejemplo, el debate sobre si la percepción o el testimonio pueden ser considerados conocimiento no es aquí relevante. Qué criterios se deben cumplir para que una proposición sea considerada conocimiento científico, sí es relevante.

---

para obtener los resultados esperados (ya sea si lo cumplen satisfactoriamente o no). En el contexto de esta sección, 'modelo' se utiliza en un sentido más general. Se trata de una representación simplificada del sistema o del fenómeno con la que se interactúa para deducir información sobre estos sin tener que lidiar con su nivel de complejidad.

Desde esta perspectiva, entonces, la definición de conocimiento con la que se trabajará, será la siguiente:

(I) *S* conoce que *p* si y solo si *S* tiene una creencia segura de que *p*, y *p* es verdadera.

S es un agente cognoscente y *p* es una proposición. Una creencia segura es aquella para obtener la cual, *S* no pudo haber sido llevado fácilmente a formar una creencia falsa (Khalifa, 2017, p. 12). Se puede observar de primera entrada que (I) respeta la forma tradicional de la epistemología. Es una variación de la versión tripartita originada en Platón, pero continuada hasta Gettier (1963), del conocimiento como creencia verdadera y justificada. Nótese, además, que no incluye una de esas tres características: la justificación. Sin embargo, de alguna manera viene incluida como uno de los potenciales motivos por los que una creencia verdadera podría ser *segura*. Este último elemento funciona como una vía para evitar los problemas referentes a la suerte epistémica que surgen del mismo Gettier (1963): en sus contraejemplos clásicos, mientras que sí se puede admitir que se trata de creencias verdaderas y justificadas (aunque por suerte epistémica), también es fácil ver que el proceso de formación de esas justificaciones pudo haber llevado *muy fácilmente* a Smith —personaje utilizado junto con Jones en los contraejemplos propuestos por Gettier en el artículo aludido— a tener una creencia falsa, dado que hubiera bastado con que él mismo no tuviera exactamente diez monedas en su bolsillo para que "el hombre que va a obtener el trabajo tiene diez monedas en su bolsillo" hubiera sido falsa, lo cual es muy plausible puesto que no era consciente de cuántas monedas del todo tenía en su propio bolsillo, mientras que sí era muy consciente del número de monedas en el bolsillo de Jones.

Se desprenden un par de observaciones de esta definición. En primer lugar, se podría considerar una proposición *q* sobre la cual S no tiene una creencia segura, pero que no es excluida por *p*. En este caso, por definición, S no conoce que *q*, pero potencialmente podría deducir el conocimiento de *q* del hecho de que *p*. La certeza de *p* sería el criterio (o uno de los criterios) de seguridad de que *q* es verdadera. En segundo lugar, se podría pensar que el problema de esta definición radica en la ambigüedad de cuándo se considera que algo pudo llevar fácilmente a S al error y cuándo no. Ante esta duda, se presenta la definición como relativa a una finalidad o contexto específico: aquello que se considera como que difícilmente lleva al error en un contexto cotidiano podría muy fácilmente llevar al error en un contexto científico. La seguridad y la facilidad en esta definición se adaptan al contexto para el que se estén aplicando: no es lo mismo cuando se habla de conocimiento en contextos legales, políticos, religiosos o éticos, que en contextos estrictamente

científicos. En cada uno de estos campos, los criterios de "facilidad" serían distintos. Sin embargo, la utilidad de la definición dada radica en que al menos provee de un marco general al que solo se deben ajustar los parámetros de la "cercanía a la producción de error" para que funcione adecuadamente.

Partiendo de esta definición de conocimiento, el subtipo llamado conocimiento *científico* se podría determinar de la siguiente manera (Khalifa, 2017)[35]:

(II) *S* conoce científicamente que *p* si y solo si la seguridad de la creencia de que *p* surge de una explicación científica *q* de *p*.

El elemento que debe ser aclarado ahora, dada esta definición, es el de 'explicación científica'. Pero para llegar a este compuesto, debemos dividirlo nuevamente en partes. Empezando por la definición de *explicación* (correcta) en general.

(III) "Se dice que *q* explica correctamente *p* si y solo si:
(1) *p* es (aproximadamente) verdadero
(2) *q* hace una diferencia para *p*
(3) *q* satisface ciertos requerimientos ontológicos (razonables)
(4) *q* satisface constreñimientos locales apropiados" (Khalifa, 2017, p. 7).

La condición (1) evita que se pueda explicar correctamente un hecho falso, lo cual sería confuso e innecesariamente contraintuitivo. Por ejemplo, una explicación que muestre por qué la Tierra es plana. De entrada, la explicación no puede ser correcta porque *p* no sería, ni siquiera aproximadamente, verdadera. Al mismo tiempo, añade el término 'aproximadamente' de modo que *p* no tenga que corresponder de forma absoluta con los hechos —objetivo cuya concreción parece poco plausible—, sino que basta con una aproximación razonable al sistema modelado.

La condición (2) introduce el elemento de lo contrafáctico en la definición de la explicación: *q* hace una diferencia para *p* si, de no darse *q*, tampoco se pudiera dar *p*. Por ejemplo, para explicar

[35] De hecho, Khalifa (2017) no da las definiciones que se plantean en (I) y (II). Aquí se parte de la definición que da el autor para el "conocimiento científico de que *q explica por qué p*" (p. 12). Al ser la definición de un definiendo más complejo que los presentados aquí, y al no ofrecer las bases conceptuales (porque no son necesitadas en el contexto en que escribe el autor), se ha optado por hacer ingeniería inversa para obtener (I) y (II). De modo que la referencia que se hace a Khalifa (2017) en (I) y (II) no implica que él suscriba esas definiciones necesariamente, nada más que estas se obtuvieron a partir de su propuesta. Las que se ofrecen más adelante sí fueron directamente propuestas por él.

el movimiento de la Luna alrededor de la Tierra (*p*), utilizo, entre otras, las diferentes leyes relativas a la fuerza gravitacional (*q*) —las newtonianas son un buen ejemplo, puesto que muestran la importancia de la condición (1): la relatividad general explica aún mejor y con más detalle el movimiento de la Luna alrededor de la Tierra, pero las leyes de la gravedad de Newton son una aproximación bastante razonable y adecuada—. En este caso, las leyes relativas a la fuerza gravitacional (newtonianas) hacen una diferencia para el movimiento de la Luna, puesto que, sin ellas, aquel probablemente seguiría patrones de muy distinta naturaleza. Nuevamente, se podría decir que ellas no son el caso, y que es la relatividad general la que es verdadera. Mientras que esto es correcto, hay al menos una correlación importante entre esta y aquellas en el contexto específico de la relación Tierra-Luna, por lo que sigue siendo verdadero que, de no ser ciertas para este caso esas leyes newtonianas, el comportamiento sería bastante diferente y, probablemente, la relatividad general tampoco podría explicarlo. Esto muestra una fortaleza de la definición dada de explicación, y es que responde al hecho de que en la ciencia no se suelen eliminar por completo modelos que siguen siendo útiles en la práctica. Todavía se siguen efectuando muchos cálculos con las fórmulas newtonianas de la gravedad porque todavía siguen ofreciendo una aproximación bastante certera y con mucho menos consumo de recurso computacional que si se calculara todo a través de la relatividad general. El intercambio entre una y otra es razonable, y la definición dada arriba no excluye esta posibilidad.

El elemento contrafáctico introducido en (2) es de suma importancia para el tema que aquí se desarrolla. De hecho, múltiples modelos de IML tienen como premisa particular el uso de 'explicaciones contrafácticas' (Aıvodji *et al*., 2020; Dandl *et al*., 2020; Slack *et al*., 2021), en lugar de las 'explicaciones no-contrafácticas' que ofrecen otros como LIME, SHAP y el método de Anclas. Nótese que, desde la perspectiva de la definición dada más arriba y, específicamente, por (2), hablar de "explicaciones contrafácticas" es redundante, al mismo tiempo que hablar de "explicaciones no-contrafácticas" es contradictorio. Es decir, LIME, por ejemplo, no estaría produciendo explicaciones del todo, a menos que sus resultados pudieran ser interpretados contrafácticamente. Sobre la importancia de lo contrafáctico para las explicaciones, se hablará posteriormente.

El elemento (3) no es tan relevante para esta investigación. Está ahí para que la definición sea adaptable tanto a posturas de realismo científico, como a posturas de antirrealismo científico. En caso de realismo científico, se esperaría que *q*, del mismo modo que *p* por (1), sea verdadera (aproximadamente). En caso de antirrealismo, para aquello que se afirme sobre entidades no

observables no haría falta que *q* fuera verdadera para que explique correctamente *p* (Khalifa, 2017, p. 7). Tiene la función, por lo tanto, de satisfacer ambas posturas según criterios requeridos por cada una.

Por último, el elemento (4) admite cierto nivel de contextualismo similar al que permite el método de los niveles de abstracción que se comentó más arriba: la corrección de una explicación depende en gran medida de la disciplina y el contexto en que se dé. Hay diferentes formas de explicar, pero no todas aplican para todo caso. (4) es importante porque nos termina de dar la definición general de explicación. Es decir, brinda el último criterio mediante el cual se juzga si algo es una explicación correcta o no. Pero además permite precisar conceptualmente la pregunta que se propuso con anterioridad: ¿qué constreñimientos locales determinan a la explicación específicamente *científica*?

Cabría preguntarse si, por ejemplo, la causalidad (no ya solo lo contrafáctico) es una condición necesaria para la explicación científica, o si existe también la posibilidad de explicaciones de otra clase. La posición más común, sin duda, es aquella según la cual la causalidad sí es uno de esos constreñimientos locales, aunque se pueda representar de diferentes formas. Pero también hay muchas posturas hoy en día que defienden la existencia de las explicaciones científicas no-causales al lado de las causales (Reutlinger, 2016). De hecho, todavía está activo el debate sobre este tema desde diversos frentes. Entre quienes admiten la existencia de explicaciones científicas causales y de las no causales, se encuentran los que proponen que ambas formas de explicación requieren de teorías particulares para cada una de ellas. Es decir, la causal y la no-causal son explicaciones de naturalezas totalmente distintas. Esta postura se conoce como *pluralismo* de la explicación. Por otra parte, están quienes defienden lo contrario: las explicaciones causales y las no-causales son explicativas por uno o varios factores que tienen en común. A estos últimos se les llama *monistas* de la explicación (‘monismo’ aquí no significa que exista *un solo tipo* de explicación, sino que *todos los tipos* de explicación se subsumen bajo uno o varios criterios compartidos) (Reutlinger, 2016, p. 6).

Ambas posturas, tanto la pluralista como la monista, tienen particularidades que cabe revisar con un poco más de profundidad para formar un concepto más claro de explicación científica. Una de las formas de justificar el pluralismo de la explicación es aquella según la cual las causales y las no-causales se distinguen entre sí porque las primeras dependen de cuestiones “ónticas”, mientras que las segundas dependen de factores “modales” (Lange, 2013; Salmon, 1984).

La explicación que se requiere para mostrar por qué una persona no puede repartir 23 fresas enteras equitativamente entre 3 personas, no es causal en la medida en que es independiente de lo óntico (y de las leyes naturales causales). Es más bien no-causal en la medida en que depende de un factor meramente modal (y matemático): 23 no es divisible entre 3. Es *imposible* dividir 23 entre 3 con un resultado entero.

Los monistas, por otro lado, afirman que hay una característica común entre las explicaciones causales y las no-causales, incluso si se acepta la diferencia entre la dependencia óntica y la modal (Reutlinger, 2016). El concepto no es ajeno a esta investigación: lo contrafáctico[36]. Uno de sus mayores promotores, Woodward (2003), lo define así: "Una explicación debe ser tal que nos permita ver qué clase de diferencia hubiera hecho para el *explanandum* si los factores explicados en el *explanans* hubieran sido diferentes de varias maneras posibles." (p. 11). En la definición de lo experimental, también se define la causalidad que se busca en términos de lo contrafáctico.

Para ligar esto al problema incumbente, ya existen varios modelos de IML que consideran lo contrafáctico como factor importante para las explicaciones producidas. Incluso algunas se han hecho con otros modelos agnósticos como base, a los que se les ha acoplado un mecanismo para obtener de sus resultados no contrafácticos, ahora sí, una explicación contrafáctica. Un ejemplo de modelo de interpretación contrafáctico es DiCE (*Diverse Counterfactual Explanation*) (Mothilal *et al*., 2020), que da explicaciones mediante acciones contrafácticas que sean factibles y diversas. Esto último se debe a otro factor importante de lo contrafáctico: para ser explicativo, debe estar tan cercano a lo que se pretende explicar como sea posible, y ser algo posible. Es decir, para explicar por qué en la Tierra una pluma cae más lento que un ladrillo de 2kg, si utilizáramos como contrafáctico el caso en el que no hubiera aire y, por lo tanto, resistencia aérea del todo, entenderíamos por qué ambas caen a distintas velocidades: es la resistencia de ese aire que sí existe lo que las hace caer a velocidades distintas. Por otro lado, si en lugar de eso utilizáramos como contrafáctico el caso de que la Tierra ejerciera una fuerza gravitacional sobre todas las cosas dentro de su atmósfera tal que la resistencia del aire produjera una diferencia en la velocidad apenas

[36] La postura de Khalifa (2017) parece ser, en este sentido, monista, aunque en el artículo citado se considera a sí mismo pluralista (p. 8). Podría ser que, cuando lo hace, se refiere únicamente a la condición (4) de las que revisamos anteriormente, y no a su postura en general. O, tal vez, aunque considere que se comparte lo contrafáctico, puede que también considere que eso no es suficiente para explicar ambas, sino que son distintas por *otras* características esencialmente distintas.

significativa, al estar más alejado del caso que nos interesa (por ser más improbable y por requerir de más asunciones) podría resultar mucho menos explicativo: ¿Entonces es por la cantidad de fuerza gravitacional que caen a velocidades distintas? No es el caso, pero no se podría culpar a alguien por asumirlo con esa explicación. Entonces, cuanto más factible y cercano al *explanans* esté el caso contrafáctico, tanto más explicativo es. Aquí, algo está más cercano al *explanans* cuanto menos y menores sean las cosas que se deban asumir. En el caso de IML, una explicación contrafáctica seguiría estos mismos criterios: cuanto más cercano al modelo que busca hacer interpretable, tanto más explicativo será. Por ejemplo: si no estuvieran los pixeles en los que se encuentran la cola y la forma del hocico de un perro, no se habría clasificado la imagen como de un perro.

Existe una postura más, que es aquella según la cual realmente solo existe *un* tipo de explicación, y es la causal (Reutlinger, 2016). Las explicaciones pueden ser no-causales solo en apariencia: si se analizan lo suficiente, van a revelar algún elemento causal del que dependen. Esta postura, no obstante, requiere cada vez de más formas de racionalizar explicaciones "en apariencia" no causales conforme van apareciendo o analizándose nuevos potenciales contraejemplos, por lo que se asumirá aquí como poco plausible.

En lo que sigue, se asumirá la postura monista sobre la explicación específicamente científica. La definición de explicación en general (III) es, de suyo, monista también, en tanto que requiere de lo contrafáctico por la condición (2) y, al ser la científica un subconjunto de (III), donde los constreñimientos locales apropiados son los científicos, entonces también debe participar de esa condición. Nótese que esto no significa necesariamente que la explicación científica pueda ser no-causal. La contrafactualidad es necesaria por (2), pero los subconjutos de causalidad y de no-causalidad están separados dentro del conjunto de contrafactualidad. Sin embargo, por la cantidad de ejemplos de explicaciones científicas no-causales (Reutlinger, 2016, p. 4), parece razonable aceptar que la causalidad no es esencial a estas[37].

Cabe preguntarse aquí qué lugar ocuparían las explicaciones que se basan en relevancia estadística, si son causales o no-causales. Esta pregunta es de importancia para la presente investigación por el simple hecho de que muchos de los modelos de ML explican mediante relevancia estadística y, al mismo tiempo, muchos modelos de IML que buscan explicar a aquellos, lo hacen también mediante relevancia estadística (véanse los modelos agnósticos presentados en

[37] Por ejemplo, explicaciones matemáticas, topológicas, teoréticas de grafos, geométricas, abstractas, estructurales y estadísticas.

el capítulo precedente). Inicialmente estas explicaciones eran concebidas como estrictamente causales (Salmon, 1984), ya que asumían que las explicaciones en general debían todas ser causales y que las relaciones causales podían ser capturadas en su totalidad por la relevancia estadística (Woodward & Ross, 2021). Lo primero no calzaría con la posición que se acaba de adoptar en esta investigación (hay explicaciones científicas no-causales), y lo segundo ha sido probado como falso. La prueba de esta falsedad funcionará como premisa para determinar que las explicaciones de relevancia causal pueden en principio ser científicas, pero que muchas veces no lo son en la práctica porque no satisfacen con claridad el requerimiento (2) de contrafactualidad.

Piénsese en la definición estadística de causalidad que propuso Reichenbach (1956), conocida como el *principio de causa común*: Al haber dos variables estadísticamente dependientes entre sí, X e Y, debe existir una tercera variable Z que funge como influencia causal sobre ambas. De este modo, si están condicionadas X e Y por Z, se vuelven independientes (lo que las hacía dependientes era Z) (Schölkopf, 2019, pp. 3-4). Existe el caso en el que la variable Z puede ser igual a una de las otras dos. Es decir, donde el factor común de X e Y es, por ejemplo, X como tal, lo que significa que Y depende de X porque esta última ejerce causalidad sobre aquella.

La relevancia estadística se puede expresar de la siguiente manera:

(IV) Una variable C es relevante estadísticamente cuando $P(B \mid A.C) \neq P(B/ A \wedge \sim C)$ (Woodward y Ross, 2021)

Es decir, C es relevante cuando la probabilidad de que se dé el evento B, dado el evento A, acompañado de la variable C, no es igual a la probabilidad de que se dé el evento B, dado el evento A sin la variable C. Por ejemplo, la probabilidad de que una persona esté embarazada, dado el hecho de que es un hombre, y este acompañado de que toma pastillas anticonceptivas, es exactamente la misma que la probabilidad de que una persona esté embarazada, dado el hecho de que esa persona es un hombre (y la probabilidad es 0). El hecho de tomar anticonceptivos *no es relevante* estadísticamente. Por otro lado, la probabilidad de que una persona esté embarazada, dado el hecho de que es mujer, y este acompañado de que toma anticonceptivos, es significativamente distinta de la probabilidad de que una persona esté embarazada, dado el hecho de que sea mujer (por sí mismo, sin acompañamiento). Es decir, para este segundo caso el hecho de tomar anticonceptivos *sí* es relevante estadísticamente (Woodward & Ross, 2021):

P(Embarazo | Hombre.Toma pastillas anticonceptivas) = P(Embarazo | Hombre)

P(Embarazo | Mujer.Toma pastillas anticonceptivas) ≠ P(Embarazo | Mujer)

Una definición de este tipo permite unir los ámbitos de la causalidad y la información estadística. El problema radica en que las relaciones causales son subdeterminadas (*underdetermined*) por las relaciones de relevancia estadística.

Otro ejemplo que aporta Schölkopf (2019, p. 3) resulta muy aclarador: piénsese en la correlación existente entre la frecuencia de cigüeñas, X, y la tasa de natalidad de humanos fuera de hospitales, Y, que se ha encontrado en Brandenburgo recientemente (Hofer *et al*., 2004). Se podría asumir que la dependencia de la frecuencia de cigüeñas y la cantidad de niños nacidos fuera de hospitales se debe a que las primeras causan la segunda, tal y como cuentan muchos padres a sus hijos (X→Y, donde X=Z). Pero también sería legítimo pensar que son los niños los que atraen a las cigüeñas, incrementando su frecuencia (Y→X, donde Y=Z). El último caso posible es el de que haya otro factor que influencie tanto la frecuencia de cigüeñas, como a la tasa de natalidad humana fuera de hospitales, como podría ser el desarrollo económico, o algo relativo al clima (X←Z→Y). El problema es que, teniendo solo los datos de la correlación entre X y Y, sin asumir otros factores, es imposible distinguir entre los casos X→Y, Y→X y X←Z→Y: no hay ninguna razón que nos compela a escoger entre las tres posibilidades si dependemos únicamente de la relevancia estadística. Podemos *asumir* que hay una causa de la correlación, pero nada determina *cuál* sea específicamente la causa. Por ello, se habla de que las relaciones causales están subdeterminadas por la relevancia estadística: la última no logra determinar a las primeras.

De lo anterior se puede concluir varias cuestiones. En primer lugar, que la información que provee un modelo de relevancia estadística es siempre menor que la información que provee uno causal (Schölkopf, 2019, p. 4). Esto no es menor, ya que muchos de los modelos de ML resultan en instancias de lo primero. Valga destacar que la conclusión no implica que la relevancia estadística no produzca información del todo, pero sí que hay otros modelos preferibles que optimizan esa cantidad de información, lo cual los hace más explicativos o explicativos del todo. En segundo lugar, se puede concluir que, del hecho de que haya explicaciones no-causales en la ciencia, no se desprende necesariamente que todas sean igual de informativas. En los casos del ejemplo, y como se indicó un poco antes, la relevancia estadística entre X e Y no permitió formular un caso contrafáctico, donde la falta de X, la falta de Y o la falta de Z pudiera jugar un rol. Finalmente, que las explicaciones científicas que se obtienen mediante el uso de modelos de ML, así como las explicaciones de esos modelos ofrecidas por los de IML, deben ser abordadas y utilizadas con

prudencia, puesto que los límites de la información que realmente confieren son difusos y podrían degenerar en errores importantes —como evidencia la creciente cantidad de casos de fuga de datos por uso innecesario de modelos de ML en diferentes disciplinas (Kapoor & Narayanan, 2022), sobre lo cual se discutirá en la siguiente sección—.

Antes de terminar de hablar del concepto de explicación científica, se cometería un error de omisión si no se incluyera dentro de las posibilidades la que defiende Van Fraassen (1980): la teoría pragmática de la explicación. Lo particular de esta teoría es que, a diferencia de las otras que se han comentado, incluye dentro de su núcleo dos factores que son esenciales para las explicaciones: (1) los hechos sobre las creencias, los intereses y otros factores psicológicos de quienes dan y reciben las explicaciones y (2) el contexto en el que se da la explicación. Las otras teorías monistas, pluralistas y reduccionistas no niegan que los factores (1) y (2) tengan un peso sobre las explicaciones, pero sí consideran que son adicionales (o accidentales) a un núcleo diferente que es el que se quiere representar realmente (Woodward & Ross, 2021). Como resultado de la teoría pragmática, se puede inferir que no hay criterios objetivos que determinen una buena o mala, adecuada o inadecuada explicación: o la audiencia entiende (explicativa) o no (no explicativa).

Además, y tal vez más importante, el fin de la ciencia según Van Fraassen (1980) es producir teorías empíricamente adecuadas, entendidas como descripciones correctas de observables (no como verdades sobre no observables). Esta postura es antirrealista, según lo que se distinguió al respecto anteriormente en esta sección. Pero si este es el fin de la ciencia, entonces la explicación científica no es necesaria para cumplirlo. De hecho, el autor habla de esta como una virtud pragmática de la ciencia, en lugar de ser su meta. Es decir, la explicación es *adicional*, pero no necesaria. Esto no calza con la definición de conocimiento científico dada en (II), para la que la explicación es parte esencial de lo que hace al conocimiento *científico*.

Lo anterior hace necesario argumentar que Van Fraassen confunde el concepto de explicación con el de interpretación y el de entendimiento (*understanding*). Una explicación es necesaria para que el conocimiento sea científico, pero que esta sea interpretable y, por lo tanto, entendible, es otro asunto que sí se podría afirmar que depende esencialmente de los factores pragmáticos (1) y (2). La interpretación es, desde esta perspectiva, el medio por el que se adquiere el entendimiento —dependiendo de la interpretación que se haga de la explicación, esta se entiende o no se entiende: la interpretación es un proceso, mientras que el entendimiento es un estado—. Es importante considerar estos niveles de abstracción para lograr una buena interfaz entre los

modelos de ML y sus usuarios, cuyos contextos, creencias e intereses definitivamente van a jugar un papel sobre cómo interpretan la explicación que se les ofrece. Sin embargo, el nivel de abstracción que incumbe a esta investigación es específicamente el de la ciencia, en el que se asume un contexto constante y en el que la explicación es necesaria, pero su entendimiento no (Khalifa, 2017). No obstante, la explicación científica sí debe ser *en principio* entendible, hecho que se podría incluir dentro de los constreñimientos locales apropiados para su correctitud. Habría, además, niveles distintos de interpretabilidad y de constreñimientos locales según la disciplina científica.

Dada esta base conceptual epistemológica, se puede pasar a la siguiente sección, en la que se utilizarán estas herramientas para analizar el papel de ML y los acercamientos basados en datos y BD en la ciencia: sus ventajas, desventajas, fines y problemas, así como varios ejemplos de uso en la actualidad en disciplinas específicas.

3.2 *Machine Learning* y *Big Data* en las ciencias: promesas y problemas

Como se mencionó en el primer capítulo de esta investigación, la capacidad de recolección y manipulación de diferentes tipos de datos que se tiene hoy en día es inédita. No es poco común escuchar hablar de los datos como un recurso: no como un producto o como entes individuales y discretos, sino como algo continuo y que se obtiene, se procesa y se manipula como conjuntos. Hablar de "un dato" por sí mismo, es como hablar de "un átomo" al pensar en termodinámica. Tiene sus características propias, pero lo que interesa realmente es cómo interactúa con el resto de partes y las características que exhibe el todo en consecuencia, no simplemente como la suma discreta de sus partes. Aunado a ese creciente acceso a datos manipulables y utilizables, se encuentra la producción constante de herramientas que permiten procesar y refinar este relativamente novedoso recurso.

Estos dos factores han dado paso a un replanteamiento teórico de lo que es posible conocer científicamente. Hace dos milenios y medio, ya Aristóteles detectaba la diferencia entre las disciplinas que podían ser precisas, como la matemática, la geometría y la lógica, y las que, por su naturaleza compleja, llena de "diferencias e incertidumbres" no podían ser precisas, sino que solo podían aspirar a "explicar la verdad *grosso modo* y en bosquejo" (*Ética a Nicómaco*, I, III, 1094b11-26). Ahora, sin embargo, las ciencias sociales, a las cuales incluía entre esas cuestiones por naturaleza imprecisas, han encontrado una vía que les permite adoptar la complejidad sin el costo

de la precisión. A lo mejor ya no sea necesario hablar de la verdad "*grosso modo* y en bosquejo", sino que se pueda determinar verdades específicas y precisas, basadas en la recién adquirida capacidad de recolectar, procesar y analizar cantidades enormes de datos. Esto último es lo que promete el uso de ML y BD en las diferentes ciencias, no solo las sociales: hacer cognoscible lo que, sin ellas, era prácticamente incognoscible. Desai *et al*. lo ponen de la siguiente manera:

> Cuando se pone [la calidad de los datos] junto a las descripciones de su análisis y presentación, se revela una concepción de limitaciones humanas cuando se confronta a los datos, y con la ciencia de datos vista como *la* tarea epistémica para exceder esas limitaciones. (2022, p. 6)

Entendida la ciencia de datos como una disciplina que esencialmente conlleva el uso de herramientas computacionales. No obstante, esta nueva práctica no viene sin una cantidad considerable de problemas que, o ya se han presentado, o potencialmente podrían presentarse. A continuación, se presentarán las utilidades que se le atribuyen a ML y BD en la producción de conocimiento científico y posteriormente se evidencian algunas limitaciones.

Una utilidad para la ciencia de los modelos de ML, creados muchas veces gracias a BD, es que permiten la exploración y el descubrimiento científico mediante los patrones que evidencian. En general, se pueden pensar los datos como una muestra finita del mundo en la que, mediante ML, se pueden encontrar estructuras generalizables ("que se destilan en información") y aplicables a contextos distintos del original (Desai *et al*., 2022, p. 10). Dentro de esas estructuras y patrones, se puede buscar (explorar) y encontrar (descubrir) información que antes no se poseía. Algunos de los descubrimientos son posibles con un modelo de ML (Cichy & Kaiser, 2019), mientras que otros requieren de la explicación del modelo de ML que produzca otro de IML (Du *et al*., 2019; Watson & Floridi, 2020; Zednik & Boelsen, 2021).

La exploración en la ciencia es el paso que se da cuando no existe una teoría convincente que englobe el fenómeno que se quiere estudiar de la cual se puedan extraer hipótesis (Cichy & Kaiser, 2019, p. 9). Se distingue, por lo tanto, de una función más obvia de ML como lo es la predicción. Un ejemplo particular es el de las DNN en las ciencias cognitivas como *modelo* de distintas funciones. En este caso lo importante no son los resultados específicos uno y otro modelo de DNN, sino la estructura de los modelos de DNN en sí misma, ya que podría ofrecer intuiciones que después podrían evolucionar en teorías completas sobre alguna función cognitiva. La ventaja de los modelos de DNN para este fin radica justamente en su complejidad inherente, lo cual permite

un amplio espacio de exploración. Para ello es necesario “jugar con los modelos, explorando cómo se comportan y familiarizándose con ellos” (Cichy & Kaiser, 2019, p. 9). De ello se obtiene la posibilidad de generar nuevas hipótesis, la posibilidad de realizar pruebas de concepto para potenciales soluciones a problemas prácticos y la posibilidad de cambiar, precisar o crear nuevos conceptos para comprender el fenómeno estudiado. Siempre es importante tener claro que, al no tratarse de teorías maduras, sino de una etapa incluso previa a ser teoría del todo, el uso de estos medios para la exploración científica no puede ser demasiado preciso, al mismo tiempo que hay que evitar llevar demasiado lejos la analogía del modelo con el sistema que modela: las características de una y otra muchas veces no son compartidas.

El caso particular de DNN en ciencias cognitivas no es sino una instancia de algo más general: el uso de modelos para la exploración, el descubrimiento y la explicación en la ciencia. Coincidentemente, en esta investigación han convergido tres sentidos distintos de la palabra ‘modelo’ (como se advirtió en la nota 1 del primer capítulo). Mientras que es comprensible que esto haya generado y vaya a generar algo de confusión, no hay más alternativa que aclarar cada uno de ellos y, cada vez que se utilice uno nuevo, distinguirlo de los otros. El mencionado en este caso tiene un sentido distinto del del contexto de ML que aparece mayoritariamente en la investigación. Tampoco es exactamente el sentido ofrecido al hablar del método de los niveles de abstracción en la sección anterior, sino una versión más específica y tangible de este. Entiéndase un modelo (en su tercer sentido) de algo como un *ejemplo hipotético* de ese algo, donde lo hipotético significa que es una descripción de un tipo de caso, no de una instancia de ese caso (Williamson, 2017). Por ejemplo, el modelo del ciclo del agua es un ejemplo hipotético del ciclo del agua en tanto que no representa una instancia localizada de ese ciclo del agua, sino el tipo de caso que es el ciclo del agua. Estos ejemplos hipotéticos son valiosos en tanto que eliminan una gran cantidad del ruido que se presenta en las instancias en que se dan esos casos (los fenómenos mismos que se modelan), lo cual hace más fácil su exploración y realizar inferencias partiendo de su descripción. La construcción de modelos debe ser cuidadosa, no obstante, ya que no es trivial el saber cuándo una abstracción o simplificación elegida para la representación del fenómeno está removiendo una característica que es significativa o hasta esencial, al tiempo que debe haber consciencia sobre si una conclusión obtenida del modelo se debe a un factor que está en él por conveniencia matemática o de algún otro tipo, pero no en el fenómeno modelado (Williamson, 2017).

La estructura de DNN como modelo no es el único caso en el que se ha utilizado para obtener conclusiones o intuiciones sobre un fenómeno que se le asemeja. Incluso en física teórica se ha utilizado DNN, no como herramienta para predecir, sino como modelo para comprender mejor diferentes objetos de estudio mediante la exploración que aporta (Tanaka *et al*., 2021). Por ejemplo, para entender la evolución del tiempo de sistemas dinámicos hamiltonianos, en lugar de pensar en las DNNs como una forma de expresar una variedad de funciones no lineales, se puede pensar en ellas como ondas de información que se propagan entre las capas de la red. Pensado de esta forma, su comportamiento es análogo al de los sistemas dinámicos que se quieren explorar y los modela. De la misma forma, se han utilizado los diagramas de redes neuronales para explorar con mayor detalle los sistemas cuánticos de muchos cuerpos, debido a que estos se teorizan mediante el uso de una red de tensores para aproximarse a una función de onda que, en estructura, es muy similar a aquellos diagramas (Tanaka *et al*., 2021).

Además del uso que muestran estos ejemplos de las estructuras complejas ML como modelo —en el sentido de Williamson (2017)— para la exploración y el descubrimiento, también es posible utilizar los resultados de IML con los mismos fines. Ya se había comentado con más detalle en el capítulo anterior otras de las finalidades de IML: la depuración y la validación. La primera hace referencia a la capacidad que brindan de localizar problemas en modelos complejos y, por lo tanto, de solucionarlos. La segunda se refiere a la cuestión ética, en tanto que permiten revisar si, a pesar de que el modelo funcione bien en términos técnicos, está fallando por depender de sesgos externos que pueden provenir de quien lo entrenó, o de los datos mismos con que se entrenó sin que la encargada lo notara en primera instancia. Pero se había mencionado también una tercera finalidad que no se desarrolló en detalle para hacerlo aquí: el descubrimiento científico.

A través de las explicaciones que se derivan de IML, se pueden obtener intuiciones que antes no eran posibles por la opacidad epistémica del modelo explicado (Du *et al*., 2019; Watson & Floridi, 2020; Zednik & Boelsen, 2021). Se podría argumentar que esto también es posible únicamente sobre la base de las predicciones mismas que se hacen mediante ML complejo, y no sería incorrecto. Sin embargo, estas hipótesis estarían sujetas a la duda que produce lo opaco del funcionamiento del modelo, mientras que las que dependen de la explicación de IML parecerían (inicialmente) exentas de este problema. En otras palabras, con IML se podría extraer información de los datos que no estaría presente en los *outputs* de los modelos de ML, sino en la configuración y el funcionamiento mismo del modelo que produce esos *outputs*.

La aplicación de IML en química ilustra lo dicho muy bien. En esta disciplina, al igual que en prácticamente todas las ramas de la ciencia en la actualidad, ha ido en crecimiento el uso de las herramientas que provee ML, las cuales han vuelto plausible hacer predicciones muy precisas sobre ciertos fenómenos que antes no lo permitían (Dybowski, 2020). Por ejemplo, para acelerar el proceso de descubrimiento de drogas, se utilizan ANN para predecir si un compuesto químico interactuaría con cierta partícula objetivo (alguna biomolécula asociada con una enfermedad). En particular, se utiliza una red neuronal convolucional llamada AtomNet, que ha sido entrenada con un rango enorme de interacciones entre compuestos proteínicos, para llegar a esas predicciones con alta precisión. Esto acelera muy significativamente el proceso de crear drogas comerciales y útiles, al tiempo que reduce su costo.

El problema es que, si se contentaran con esas predicciones, el uso de datos y de recursos para el entrenamiento de la red neuronal sería sumamente ineficiente, además de que desaprovecharía la oportunidad de obtener intuiciones más significativas y que podrían ahorrar trabajo y formación de hipótesis a futuro. Por ello, se han utilizado diversos métodos de IML —tanto agnósticos, como específicos para el modelo—, entre los cuales está LIME, para obtener resultados más allá de las predicciones. Se pudo descubrir mediante gráficos de dependencia parcial, por ejemplo, las subestructuras químicas que permiten diferenciar entre compuestos tóxicos y no tóxicos (Dybowski, 2020, p. 20916). Esta información no era explícita partiendo únicamente de las predicciones, sin utilizar IML sobre ellas, pero constituye una instancia de conocimiento científico adquirido (descubierto) gracias a las explicaciones del modelo. Situaciones análogas se han presentado también en el desarrollo de hipótesis en física (Greitemann *et al*., 2019; Iwasaki *et al*., 2019; Kapteyn & Willcox, 2020).

Por casos como el anterior, se ha argumentado que la importancia de IML en el contexto de exploración científica es mayor que la de los modelos predictivos de ML (Zednik & Boelsen, 2021). Sin estas explicaciones la extracción de información desde los conjuntos de datos sería muy ineficiente, permitiría menos intuiciones y, en general, reduciría el alcance científico de la investigación basada en datos. Un concepto útil en este marco es el de “incertidumbre de vínculo” (Sullivan, 2019). Esta aparece cuando no hay evidencia empírica que vincule las causas *posibles* que se identifican en un modelo, con las causas *reales*. Esta situación, en la que el modelo no provee mediante sus meras predicciones clara certidumbre sobre el vínculo entre él y el fenómeno real,

evidencia la necesidad de reducir esa incertidumbre, lo cual es justamente uno de los propósitos de IML.

Se puede abstraer aún más lo anterior: de hecho vincular las causas *posibles* en los modelos (en el sentido de Williamson (2017)) con las causas *reales* del fenómeno es exactamente una de las funciones de las explicaciones científicas causales en general. Las explicaciones científicas no-causales, al ser modales en lugar de ónticas, como se había determinado anteriormente, no sufren del problema de la incertidumbre de vínculo, ya que no hay contraste con algo empírico, sino que dependen de constreñimientos modales lógico-matemáticos. Esta abstracción es significativa, ya que apoya el argumento de la importancia de IML al utilizar ML en la práctica científica al mostrar que, al aportar una explicación aproximadamente correcta del funcionamiento del modelo de ML, está reduciendo no trivialmente la incertidumbre de vínculo entre él y el fenómeno que está modelando. O, en otras palabras más directas, IML da a los modelos epistémicamente opacos el estatuto de "científicos".

Hay quienes defienden que, además del poder exploratorio y generador de hipótesis o inferencias, los modelos pueden brindar explicaciones científicas por sí mismos (Bokulich, 2011). Si volvemos a las consideraciones sobre la explicación científica que se dieron con anterioridad, recordaremos que lo contrafáctico —aquello que, de estar ausente, produciría una diferencia— es parte esencial de esta. Un modelo adecuado permite interactuar con las partes que lo conforman y visualizar con mayor claridad la diferencia que hacen según cambios cuantificables, lo cual se puede extender (teniendo en cuenta las limitaciones expresadas más arriba) al fenómeno que se modela.

No obstante, si llevamos este argumento a la particularidad de los modelos de ML, podríamos diferenciar entre los casos en que es verdadero —los modelos de ML que son de suyo interpretables, como un árbol de decisiones simple o una regresión lineal—, de aquellos en los que no se cumple —cuando los modelos de ML son opacos por su complejidad—. Estos últimos son modelos con los que no se puede jugar, explorar cómo se comportan y familiarizarse con ellos (para retomar la cita de Cichy & Kaiser (2019, p. 9) ofrecida más arriba) en la misma medida que con modelos transparentes. Después de todo, la ventaja que muchos autores coinciden en darle a los modelos (en el sentido de Williamson (2017)) como herramienta para exploración científica, es el hecho de que son *simplificaciones* del sistema que modelan. Pero lo que caracteriza a los modelos opacos es que son tales por su *complejidad*.

Por lo anterior, no se puede negar que la ciencia de datos ha abierto una nueva vía a las diferentes disciplinas científicas para realizar sus prácticas de manera más eficiente. Eso sí, debe ir acompañada del conocimiento de sus limitaciones epistemológicas y de las herramientas necesarias para que no se confunda el *método* con la *teoría*. Podría parecer como una advertencia condescendiente e innecesaria, de no ser porque esta confusión y ese uso imprudente de los modelos complejos de ML es lo que está ocurriendo de manera acrítica en la práctica actualmente, al punto que incluso se habla de una crisis de la replicación en la ciencia basada en ML (Kapoor & Narayanan, 2022). Después de analizar 20 artículos en 17 campos distintos en los que se han encontrado errores debidos al fenómeno conocido como fuga de datos, Kapoor y Narayanan (2022) determinaron que 329 artículos que dependían de esos 20 iniciales fueron afectados por los errores. Esto se puede categorizar como una crisis de la replicación. Especialmente preocupante es el hecho de que las publicaciones no replicables tienden a ser citadas con mayor frecuencia que las replicables (Serra-García & Gneezy, 2021), del mismo modo que la información falsa tiende a propagarse con mayor velocidad que la información verdadera en redes sociales (Vosoughi *et al*., 2018). Lo que ocurre es que muchas disciplinas científicas están adoptando la modelación predictiva que ofrece ML por ser un método novedoso y que promete muchas nuevas rutas de exploración. Sin embargo, al utilizar estos métodos acríticamente, no son conscientes de sus limitaciones y potenciales problemas, como la mencionada fuga de datos.

La fuga de datos se da cuando el conjunto de datos utilizados para entrenar un modelo de ML contiene información sobre aquello que se quiere predecir, la cual no estará disponible cuando se hagan predicciones realmente. Esto lleva, por lo general, a resultados demasiado optimistas sobre la precisión del modelo entrenado, lo cual, a su vez, tiene como resultado reacciones demasiado optimistas de quien los lee y hace más pronta su propagación, como evidencia la investigación de Kapoor y Narayanan (2022).

La verdad es que en muchos casos en que se presentan errores de replicación, el uso de un modelo de suyo interpretable de ML, uno que no introduzca los problemas que conlleva la opacidad epistémica, es la alternativa más razonable y epistémicamente virtuosa (Rudin, 2019). Ya se había expresado esta idea en el capítulo anterior cuando se concluyó que los modelos de IML no eran suficientes para solucionar el problema de la incompletitud de los modelos de ML complejos. Puede ser que los mencionados conjuntos de Rashomon no se sigan lógicamente, pero en la práctica (científica) son muchas veces el caso: para lograr las predicciones que se desean en las diferentes

disciplinas científicas, es común que no exista ese intercambio entre precisión e interpretabilidad de que tanto se habla, de modo que suele ser posible encontrar un modelo de suyo interpretable de ML que funcione con una precisión aproximadamente igual a la de otro modelo más complejo. Esta objeción no aplica para aquellos usos señalados en diferentes campos en los que lo importante no son las predicciones de un modelo complejo de ML, sino su misma estructura, su arquitectura (como se vio en los casos de física en los que se analogan fenómenos con los comportamientos de una red neuronal multicapa, o funciones similares).

Se han revisado varios usos que se le da en la práctica a los modelos complejos de ML. Junto a esto, se ha mostrado que esos usos, en tanto que herramientas científicas para llegar a conclusiones particulares, son legítimos, pero que son potencialmente ineficientes si no consideran un acercamiento explicativo que haga más transparente el modelo. Pero, además de ineficientes, muchas veces pecan de complejizar mucho más de lo necesario la forma de alcanzar sus fines, lo cual puede tener repercusiones que afecten a las disciplinas en las que se encuentran a través de una crisis de replicación que surge por la aplicación falente y, peor aún, innecesaria de tales modelos. Armada la investigación de los ejemplos que se vieron en esta sección, se puede pasar a la siguiente (y última), en la que se concluirá si el uso de ML complejo y BD han cambiado los fundamentos de la producción de conocimiento científico.

3.3 Sin teoría, sin explicación, sin ciencia

La ciencia basada en datos es la contraparte de la ciencia basada en hipótesis. No son conjuntos excluyentes, no obstante, en tanto que la última no implica que no se utilicen datos, y la primera no implica la imposibilidad de formar hipótesis. Esta relación entre contrarios significa el énfasis distinto de cada una. Por varias décadas, prácticamente desde los aportes más significativos de Popper a la filosofía de la ciencia, la imagen principal que se tiene del método de producción de conocimiento científico ha sido el del planteamiento de una hipótesis, seguida de su puesta en prueba (Elliott *et al*., 2016). BD y ML han dado paso a un cambio en esta imagen, de modo que se podría hablar de que la ciencia, al tender a estar basada en datos actualmente, no depende de ya de las hipótesis que se puedan extraer de una teoría previa, sino que son los datos empíricos recolectados y procesados los que van a determinar la dirección de la investigación (Leonelli, 2020).

No es extraño encontrar en la literatura sobre el apogeo de BD en la ciencia la postura de que este conlleva un "fin de la teoría" (Tansley & Tolle, 2009; Mazzocchi, 2015). Los datos dejan de ser un medio para alcanzar aquella, y se tornan en el objetivo mismo de la ciencia, como si se tratara de una nueva y radical versión del empirismo resurgiendo. Parece implausible negar que la ciencia ha sufrido algún cambio en la práctica debido al acceso a este recurso que hasta recientemente se pudo empezar a aprovechar. Pero, ¿es este cambio fundamental? ¿Podría hablarse de un cambio de paradigma completo de la práctica científica?

A lo largo de esta investigación se ha mostrado cómo los modelos complejos de ML, BD e incluso IML pueden contribuir a la ya establecida práctica de la ciencia. Sin embargo, ha quedado claro que ellos no comprenden una excepción a los criterios de la producción de conocimiento. En otras palabras, sería iluso negar que estas herramientas han provocado un cambio cuantitativo y cualitativo en la ciencia. De hecho, han posibilitado investigaciones que antes, al modo aristotélico, no tenían manera de dar fruto razonable por su complejidad. Pero el salto de aceptar esto a proponer un fin de la ciencia como la conocemos no parece justificado del todo. Dicho de otro modo, la revolución que proveen las herramientas aquí analizadas, no afectan a la estructura de la ciencia como un todo, sino solo parte (pequeña) de ella, debido a que posibilita análisis que antes eran técnicamente imposibles: ya no debe restringirse a relaciones lineales, sino que se pueden modelar relaciones no-lineales, más complejas y que permiten mayor detalle interpretativo y explicativo.

Existen nuevos problemas, aparecen nuevos métodos para resolverlos, y nada de esto elimina la necesidad de que las proposiciones científicas —o incluso los modelos científicos, en caso de que no se asuma que todo conocimiento científico deba ser proposicional— deban ser explicativas de aquello sobre lo que teorizan. Si se entiende el conocimiento científico como se hizo aquí en (II), la capacidad de explicar no es, como quería Van Fraassen (1980), una virtud epistémica adicional de la ciencia, sino que es una característica fundamental de la misma. Así, por más que al pasar cantidades ingentes de datos por algoritmos de ML complejos se obtengan resultados muy precisos sobre algún objetivo, en ese proceso no hay producción de conocimiento científico del todo: de nuevo, se confunde el *método* (y sus herramientas) con la *teoría*.

Este punto se demostró con claridad en la sección anterior: depender únicamente de la precisión de los resultados de un modelo complejo de ML es hacer ciencia ineficiente, que desperdicia toda la información teórica que podríamos obtener del recurso de los datos al no intentar hacer un poco más transparente su caja negra. Los modelos de DL encuentran

correlaciones, no lineales, pero que no están determinadas desde una perspectiva explicativa. Por eso, si se depende solo de la precisión de los resultados, se ponen en peligro numerosísimas situaciones: usar un modelo sobreajustado, de lo cual no se da cuenta el investigador por no interpretar y depurar el modelo; usar un modelo sesgado, tanto por preferencias personales de la investigadora, como por simples sesgos estadísticos no visibles para la misma; haber utilizado un modelo con fuga de datos; causar una crisis de replicación por todas las razones anteriores; y, en general, dejar olvidadas intuiciones importantes para poder llegar a nuevas hipótesis y nuevos descubrimientos científicos por no hacer transparente por qué el modelo obtiene los resultados que obtiene.

De nuevo, se tiene conocimiento científico cuando está acompañado de una explicación científica, y la ciencia de datos tiene todo un campo dedicado a poder producir esas explicaciones para hacer mucho más eficientes y confiables los modelos complejos de ML y su procesamiento de BD. Para quien teme que vaya a ocurrir el fin de la teoría, se ofrece el siguiente experimento mental: asúmase que en efecto la ciencia llegue a ese empirismo extremo del que se habló más arriba, donde la teorización y la explicación ya solo tienen un papel secundario y poco significativo. Piénsese que la ciencia ahora consiste en procesar cantidades gigantescas de datos recolectados automáticamente mediante modelos cada vez más complejos y, por ende, más opacos de ML. En este mundo una científica podría cumplir roles limitados, en la medida en que solo podría aportar a elegir la arquitectura correcta de ML, etiquetar los datos recolectados, en caso de que no sea una arquitectura no-supervisada, y poco más que ello: los resultados de las predicciones de los modelos son todo lo que se necesita. Las máquinas estarían prácticamente haciendo ciencia por nosotros. En este escenario tenebroso, esa científica no sería realmente tal, sino que sería alguien con competencias técnicas: ¿se podría decir que ella conoce científicamente algo? ¿Se podría decir que la sociedad humana está constante y automáticamente adquiriendo conocimiento gracias a esas máquinas?

La respuesta a la que se llega si se considera todo lo que se ha analizado en los capítulos y secciones anteriores de esta investigación es negativa en ambas preguntas. ¿Cómo se diferenciaría la situación de ese mundo de ciencia automática del mundo de los seres humanos del paleolítico? A fin de cuentas, los humanos de hace más de 15000 años encontraban una cantidad ingente de datos empíricos mediante sus percepciones e interacciones con el entorno. No tenían forma de recopilarlos masivamente, y mucho menos de procesarlos en el nivel en que lo podemos hacer

ahora, en el que juegan el rol de recursos continuos, casi como la electricidad. Pero, aunque hubieran podido recolectarlos, ¿qué hubieran podido haber hecho con ellos? Poco o nada. ¿No se encuentra nuestra científica-técnica en exactamente el mismo lugar? Si no tiene ella acceso a la caja negra de los modelos de ML que utiliza, no tiene una forma de comprender por qué las máquinas llegan a las predicciones que llegan, y estas son cada vez más complejas y los datos son cada vez más, más variados, más detallados y más específicos, al punto que convergen con la realidad misma, con los fenómenos que "modelan". Cada vez la distinción entre modelo y realidad sería menor, de modo que cada vez más la científica se encontraría más cerca de la posición en la que se encontraba el humano paleolítico. Evidentemente, no se podría llamar a este modelo convergente a realidad "ciencia". El punto de la ciencia es justamente generalizar a partir de lo particular, no converger hacia lo particular (el sistema que intenta modelar), y mientras que los modelos complejos de ML no permitan esas generalizaciones, esa teorización y esas explicaciones, no se puede llamar a su producto ciencia.

Aquí influye un problema al que se aludió en el primer capítulo (nota 3), pero para el que no se contaba aún con las herramientas conceptuales necesarias para trabajarlo: la relación entre cantidad y calidad desde la perspectiva de la ciencia. El concepto de sobreajuste será de gran ayuda. Hay dos formas en las que se puede entender la relación entre la cantidad y la calidad. Se asume que con cantidad se hace referencia a los datos: cuántos datos se poseen. Este factor ha ido en exponencial aumento en los últimos años. Sin embargo, la calidad puede hacer referencia a dos cosas: a los mismos datos o a la ciencia producto de esos datos. Cuando se piensan de la primera forma, es decir, cuando se hace referencia a la cantidad y calidad de los datos mismos, es más fácil separar las dos categorías. Por un lado, está el número de datos. Por otro, su calidad general. En este caso, parece ser cierto que, cuanto más indiscriminadamente se recojan datos, tanto menor será su calidad. No hay criterio de selección ya que se suele realizar de manera automática. La calidad aumenta en la medida en que aumenta también la calidad de las herramientas que los recogen. Especialmente en el contexto de la ciencia de datos, estructurar un conjunto de datos útil y adecuado no es siempre una tarea fácil. Por otro lado, la calidad de la ciencia realizada con esos datos depende en buena medida de la calidad de los datos mismos, pero también depende de su cantidad. Es decir, la calidad de la ciencia basada en datos es una función de la cantidad y la calidad de los datos mismos. Esto implica necesariamente que la *cantidad* de los datos puede repercutir directamente también sobre la *calidad* de la ciencia, tanto para bien como para mal. Hay límites para lo anterior: si la cantidad es excesiva, puede darse el fenómeno explicado de sobreajuste,

disminuyendo la calidad de la ciencia. Si la cantidad es muy reducida, la ciencia puede no ser representativa del todo. Así, se prueba lo dicho en la nota referida.

Como evidencia la creciente crisis de replicación en las ciencias que utilizan ML erróneamente (Kapoor & Narayanan, 2022), probablemente por la emoción que generan métodos novedosos y sus potenciales resultados, pero también por negligencia e incomprensión, tanto epistémica como técnica, de cómo deben ser puestos en práctica, estas no son herramientas que se puedan utilizar indiscriminadamente. Si se sigue actuando de esa forma al intentar producir conocimiento científico, se está peligrando no cumplir con tal fin del todo. Este punto ha necesitado ser aclarado en cada dominio en particular en que se aplica. Por ejemplo, al recolectar datos no estructurados sobre biodiversidad en cantidades que se pueden etiquetar con el nombre de BD con la finalidad de producir conocimiento científico sobre los comportamientos poblacionales de especies para actuar de manera más efectiva en su conservación, muchos expertos en la materia han dejado de lado la *calidad* de esos datos (Bayraktarov *et al*., 2019). Así, a pesar de que sus intenciones son buenas en términos epistémicos y éticos, están afectando negativamente su propio objetivo. Una gran cantidad de datos y un modelo de ML preciso no implican, por sí mismos, conocimiento científico, y más bien podrían generar desinformación en la que se podrían basar acciones que perjudiquen aquello mismo que se quiere ayudar (Bayraktarov *et al*., 2019).

En parte contrario a lo dicho, existen quienes defienden, no solo que no es posible dar una explicación que haga significativamente transparente ciertos modelos con un nivel de complejidad alto, sino que además no es deseable, tanto en términos del progreso mismo de ML como campo, como del progreso de la ciencia en general gracias al beneficio que ML aporta (Zerilli *et al*., 2019). El argumento no establece que IML sea una tarea inútil, o que la transparencia y la interpretabilidad no sean aspectos importantes de considerar en la aplicación de estos modelos. Pero afirma que el estándar de explicación y transparencia que se le está pidiendo a los modelos complejos de ML para ser viables es implausible y poco realista. La premisa principal es que muchas de las decisiones humanas están cargadas de opacidad. A pesar de que sean justificadas, la justificación suele ser *post hoc* y no muy adecuada en muchos casos. Por tanto, exigir a los modelos de ML que no sean del todo opacos para poder ser utilizados es poner una barra demasiado alta e innecesaria, ya que ni los humanos la satisfacen (Zerilli *et al*., 2019).

En el segundo capítulo se habló del posible problema de coordinación entre IML y ML que evita una comunicación certera y sin ruido, el cual implica que, en efecto, la transparencia de ciertos

modelos complejos de ML nunca va a ser absoluta. Sin embargo, a lo largo del presente capítulo se ha visto que no es necesaria una transparencia absoluta para que puedan aportar en términos científicos. Similar a lo que argumentan Zerilli *et al*. (2019), aquí se defiende que el problema para la ciencia es la opacidad absoluta, pero que la transparencia absoluta no es requerida para adquirir todas las ventajas que brindan los modelos de ML a la práctica. Después de todo, lo que se requiere en la ciencia es el nivel de transparencia necesario para poder *formar una teoría*, *explorar el modelo* o *explicar un fenómeno*. Cuánta transparencia sea, dependerá del fin para el que se utiliza en primer lugar. Entonces, aunque se plantea como requisito que el modelo sea interpretable, ya sea por sí mismo o por IML, no se está pidiendo en un grado poco realista o implausible.

Dicho lo anterior, es importante resaltar que los cambios en la práctica científica hacia un método cada vez más basado en datos, aunque no son fundamentales y no implican el "fin de la teoría", sí conllevan un gran proceso de transición lleno de problemas nuevos e imprevistos, tanto técnicos, como epistemológicos, científicos, legales y éticos, muchos de los cuales se han señalado en secciones anteriores. Estos deben ir acompañados de teorización en todos los campos mencionados para garantizar eficiencia, utilidad y ética en esa transición. Más allá de estas observaciones, no existe tal cosa como ciencia sin teoría o sin explicación. Lo que existe ahora es ciencia con nuevos métodos de análisis para producir mejores teorías y mejores explicaciones, los cuales requieren, a su vez, teorías y explicaciones para ser utilizados efectivamente. El peligro del uso de ML opaco en la ciencia es cuando es indiscriminado, acrítico e innecesario, pero no es su uso en sí mismo. Como recomendación general, eso sí, se debe reforzar la investigación en IML con una claridad epistemológica mayor, donde las explicaciones que se produzcan sean, en efecto, explicaciones, y no un conjunto más de datos que podrían estar descoordinados con la fuente.

CONCLUSIÓN

Los cambios que ha permitido la ciencia de datos en la manera de producir conocimiento científico han sido profundos, y apuntan hacia un mundo en el que el conjunto de lo cognoscible es mayor que nunca. No obstante, la transición hacia ese mundo no debe ser precipitada, pues los errores éticos, epistémicos y políticos que ya se han cometido muestran la gravedad de los que potencialmente podrían seguir.

La ciencia como conjunto de disciplinas con diferentes métodos para la reducción de sesgos, como productora del conocimiento más certero, preciso y objetivo que existe sobre lo natural —en un sentido amplio—, tiene como uno de sus fines el reducir la opacidad epistémica: conocer aquello que no se conocía; mejorar el nivel de conocimiento sobre aquello que apenas era cognoscible; y, no en menor medida, hacer más comprensible el mundo para la sociedad humana y para los individuos que la conforman. Desde esta perspectiva, una ciencia que sea tan opaca como aquello que intenta hacer transparente, no tendría cabida. Este sería el caso de una ciencia centrada en los resultados, en la precisión de las predicciones y que deje de lado la teorización y la explicabilidad. Como se mostró en los capítulos anteriores, esta es la noción que muchos investigadores dedicados al tema manejan: no importa la transparencia (relativa) de la predicción, solo importa si es correcta o no. Pero detrás de una creencia tal, se esconden múltiples malinterpretaciones de lo que significa hacer ciencia y una ingenuidad y credulidad sobre la importancia de la predicción particular en sí misma. Esta ingenuidad debería servir de alerta sobre el uso inadecuado de las herramientas de ML. El problema es que la sociedad en general todavía no se aclimata epistemológicamente a los modelos de ML. Muchas personas —expertos en ML incluidos— creen, por ejemplo, que los grandes modelos lingüísticos recientes se acercan preocupantemente a la inteligencia y comprensión humanas (Mitchell y Krakauer, 2023). Más allá de que esos datos muestren posibles deficiencias en los marcos epistemológicos de los expertos en el tema, el resultado sobre el uso de modelos complejos es el mismo: si se considera que tienen una inteligencia que se aproxima a la humana, tampoco habría razón para que estén exentos de las mismas exigencias de justificación epistémica que un humano.

La explicación es, como se definió anteriormente, parte esencial del concepto de conocimiento científico. La ciencia debe ser, por lo tanto, en principio interpretable. Mientras que algunos proponen la era venidera de una ciencia sin teoría, esta investigación se posiciona en contra

de ese vaticinio. Aquello que permite hacer cognoscible la ciencia de datos y las herramientas que de ella se derivan es mucho y de mucha importancia como para minimizarlo. No obstante, siempre será requerida la teorización. Las máquinas son excelentes, en muchos casos significativamente mejores que los humanos, detectando patrones en los datos —patrones que son invisibles a los humanos—. Sin embargo, esos patrones, mientras no sean interpretados, no pasan de seguir siendo datos de los que ahora sabemos características estructurales y de relación. Para poder pasar a ser información, deben ser comprendidos, se debe teorizar sobre ellos, se deben incluir en una red conceptual que sea coherente. A los patrones se les debe asignar un significado. Esta es la labor del científico que hace ciencia basada en datos: debe saber extraer información teórica de los patrones, pero antes también debe saber hacer las preguntas adecuadas a esos datos para extraer patrones relevantes (Floridi, 2014).

Se definieron, entonces, la ciencia de datos y el concepto de ML. Se vieron algunos ejemplos de modelos con complejidades distintas, desde regresiones lineales hasta redes neuronales multicapa, y se planteó una definición precisa de BD. Luego, se determinó el problema de la opacidad epistémica y se consideraron sus límites. Se examinaron algunos modelos de IML que buscan hacer más transparentes a los modelos complejos y después se mostraron algunos problemas fundamentales de IML en general como solución a la opacidad epistémica desde la perspectiva de la ciencia. Finalmente, después de definir un marco epistemológico con los conceptos de conocimiento científico, explicación científica y explicación en general, se mostró que la ciencia debe ser explicativa por esencia y que el uso de ML adecuado que se hace en la ciencia no pierde esa característica de suyo, aunque para ello es necesario aplicar IML o hacer uso de modelos más simples e interpretables por sí mismos. Finalmente, se concluyó que hablar de un cambio de paradigma en la ciencia por la ciencia basada en datos y el uso de las herramientas de la ciencia de datos, no está realmente justificado. Sin embargo, no por ello se debe ignorar la importancia de utilizar estos nuevos métodos sin considerar cómo afectan en la práctica y qué es necesario para hacer un uso epistémicamente adecuado de ellos.

Después de examinar las razones meramente epistemológicas que causan lo que se conoce como el problema de la opacidad epistémica, tan pronunciado en el uso de modelos complejos de ML, se puede concluir que esa opacidad no es absoluta: no se trata de un problema binario en el que, o se comprende todo el modelo, o no se comprende el modelo del todo. Querer comprender cada uno de los parámetros, cada uno de los pesos y sesgos que conforman un modelo y su razón

de ser, es muy poco razonable e innecesario para poder dar una explicación científica coherente de la razón por la que ese mismo modelo produjo cierto *output*, cierta respuesta específica. Es como si, para dar una justificación epistémica de una teoría científica, fuera requisito explicar el estado de inhibición de cada sinapsis en nuestro sistema nervioso. Tal requerimiento epistémico es, no so lo imposible en la práctica, sino también poco razonable y poco útil. No obstante, como quedó claro a lo largo del escrito, la tarea de explicar y comprender las decisiones de un modelo opaco, asumida principalmente por quienes desarrollan modelos IML, es de suma importancia en la actualidad y para el futuro de estas tecnologías, así como para los usuarios, los desarrolladores y la sociedad en general. Es importante tener en cuenta sus límites, pero hacerlo no implica rechazar completamente su uso en la ciencia o para otras tareas que posibilitan.

BIBLIOGRAFÍA

Aïvodji, U., Bolot, A., & Gambs, S. (2020). Model Extraction From Counterfactual Explanations. *Arxiv Preprint Arxiv:2009.01884*, 1-12.

Alpaydin, E. (2010). *Introduction To Machine Learning* (2nd Ed). MIT Press.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On The Robustness Of Interpretability Methods *Arxiv:1806.08049.* Http://Arxiv.Org/Abs/1806.08049

Aristóteles. (2001). *Ética A Nicómaco*. Alianza Editorial.

Bahga, A., & Madisetti, V. (2016). *Big Data Science & Analytics: A Hands-On Approach*. VPT.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., Mcrae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers In Ecology And Evolution*, 6. Https://Doi.Org/10.3389/Fevo.2018.00239

Begoli, E., Bhattacharya, T., & Kusnezov, D. F. (2019). The Need For Uncertainty Quantification In Machine-Assisted Medical Decision Making. *Nature Machine Intelligence*, *1*(1), Article LA-UR-18-28661. Https://Doi.Org/10.1038/S42256-018-0004-1

Benzécri, J. P. (1983). L'avenir De L'analyse Des Données. *Behaviormetrika*, 10(14), 1–11.

Berkman, B. E., Shapiro, Z. E., Eckstein, L., & Pike, E. R. (2016). The Ethics Of Large-Scale Genomic Research. In *Ethical Reasoning In Big Data*, 53-69. Springer.

Bernstein, D. J. (2005). Understanding Brute Force. En *Workshop Record Of ECRYPT STVL Workshop On Symmetric Key Encryption*, *Estream Report*, 36, 2005-2015.

Bokulich, A. (2011). How Scientific Models Can Explain. *Synthese*, 180(1), 33–45. Https://Doi.Org/10.1007/S11229-009-9565-1

Bollobás, B. (2013). *Modern Graph Theory* (Vol. 184). Springer Science & Business Media.

Bordt, S., Finck, M., Raidl, E., & Von Luxburg, U. (2022). *Post Hoc* Explanations Fail To Achieve Their Purpose In Adversarial Contexts. *Arxiv:2201.10295v2 [Cs.LG]*. Https://Doi.Org/10.1145/3531146.3533153

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231.

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *Arxiv Preprint Arxiv:1712.04248*, 1-12.

Calafiore, G. C., & El Ghaoui, L. (2014). *Optimization Models*. Cambridge University Press.

Castelvecchi, D. (2016). Can We Open The Black Box Of AI?. *Nature News*, 538(7623), 20-23. Https://Www.Nature.Com/News/Polopoly_Fs/1.20731!/Menu/Main/Topcolumns/Topleftcolumn/Pdf/538020a.Pdf

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., & Kerren, A. (2020). The State Of The Art In Enhancing Trust In Machine Learning Models With The Use Of Visualizations. *Computer Graphics Forum*, 39(3), 713–756. Https://Doi.Org/10.1111/Cgf.14034

Christian, B. (2020). *The Alignment Problem: Machine Learning And Human Values*. WW Norton & Company.

Cichy, R., & Kaiser, D. (2019). Deep Neural Networks As Scientific Models. *Trends In Cognitive Sciences*, 23. Https://Doi.Org/10.1016/J.Tics.2019.01.009

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats Of A Replication Crisis In Empirical Computer Science. *Communications Of The ACM*, 63(8), 70–79. Https://Doi.Org/10.1145/3360311

Cox, M., & Ellsworth, D. (1997, October). Application-Controlled Demand Paging For Out-Of-Core Visualization. En *Proceedings. Visualization'97* (Cat. No. 97CB36155), 235-244.

Crawford, K., Gray, M. L., & Miltner, K. (2014). Big Data| Critiquing Big Data: Politics, Ethics, Epistemology| Special Section Introduction. *International Journal Of Communication*, 8 (10), 1663-1672.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-Objective Counterfactual Explanations. En T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, & H. Trautmann (Eds.), Parallel Problem Solving From Nature – PPSN XVI (Pp. 448–469). *Springer International Publishing*. Https://Doi.Org/10.1007/978-3-030-58112-1_31

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What Is Big Data? A Consensual Definition And A Review Of Key Research Topics. En *AIP Conference Proceedings*, 1644 (1), 97-104.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A Formal Definition Of Big Data Based On Its Essential Features. *Library Review*, 65(3), 122-135.

Dhar, V. (2013). Data Science And Prediction. *Communications Of The ACM*, 56(12), 64–73.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. *Arxiv:1702.08608 [Cs, Stat]*. Http://Arxiv.Org/Abs/1702.08608

Doshi-Velez, F., & Kim, B. (2018). Considerations For Evaluation And Generalization In Interpretable Machine Learning. En Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., & Van Gerven, M. (Eds.), *Explainable And Interpretable Models In Computer Vision And Machine Learning* (Pp. 3-18). Springer.

Du, M., Liu, N., & Hu, X. (2019). Techniques For Interpretable Machine Learning. *Communications Of The ACM*, *63*(1), 68–77. Https://Doi.Org/10.1145/3359786

Dybowski, R. (2020). Interpretable Machine Learning As A Tool For Scientific Discovery In Chemistry. *New Journal Of Chemistry*, 44(48), 20914–20920. Https://Doi.Org/10.1039/D0NJ02592E

Elliott, K. C., Cheruvelil, K. S., Montgomery, G. M., & Soranno, P. A. (2016). Conceptions Of Good Science In Our Data-Rich World. *Bioscience*, 66(10), 880–889. Https://Doi.Org/10.1093/Biosci/Biw115

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., & Van Gerven, M. (Eds.). (2018). *Explainable And Interpretable Models In Computer Vision And Machine Learning*. Springer International Publishing. Https://Doi.Org/10.1007/978-3-319-98131-4

Facebook. (2020). Facebook Reports First Quarter 2020 Results. 13. Https://S21.Q4cdn.Com/399680738/Files/Doc_News/Facebook-Reports-First-Quarter-2020-Results-2020.Pdf

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (2003). *Reasoning About Knowledge*. MIT Press.

Floridi, L. (2008). The Method Of Levels Of Abstraction. *Minds And Machines*, 18(3), 303-329.

Floridi, L. (2011). *The Philosophy Of Information*. Oxford University Press.

Floridi, L. (2014). Big Data And Information Quality. En Floridi, L. & Illari, P., *The Philosophy Of Information Quality* (303-3015). Springer International Publishing.

Floridi, L. (2019). *The Logic Of Information: A Theory Of Philosophy As Conceptual Design*. Oxford University Press.

Frické, M. (2015). Big Data And Its Epistemology. *Journal Of The Association For Information Science And Technology*, 66(4), 651-661.

Galileo, G. (1981). *El Ensayador*. Aguilar.

Gandomi, A., & Haider, M. (2015). Beyond The Hype: Big Data Concepts, Methods, And Analytics. *International Journal Of Information Management*, 35(2), 137–144.

Gantz, J., & Reinsel, D. (2012). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, And Biggest Growth In The Far East. *IDC Go-To-Market Services*. Https://Www.Cs.Princeton.Edu/Courses/Archive/Spring13/Cos598c/Idc-The-Digital-Universe-In-2020.Pdf

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis*, 23(6), 121–123.

Geweke, J. (1989). Bayesian Inference In Econometric Models Using Monte Carlo Integration. *Econometrica: Journal Of The Econometric Society*, 1317-1339.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning*. MIT Press.

Graupe, D. (2013). *Principles Of Artificial Neural Networks*. World Scientific.

Greitemann, J., Liu, K., & Pollet, L. (2019). Probing Hidden Spin Order With Interpretable Machine Learning. *Physical Review B*, 99(6), 060404. Https://Doi.Org/10.1103/Physrevb.99.060404

Hofer, T., Przyrembel, H., & Verleger, S. (2004). New Evidence For The Theory Of The Stork. *Paediatric And Perinatal Epidemiology*, 18(1), 88–92. Https://Doi.Org/10.1111/J.1365-3016.2003.00534.X

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, 2(5), 359-366.

Humphreys, P. (2009). The Philosophical Novelty Of Computer Simulation Methods. *Synthese*, 169(3), 615–626. Https://Doi.Org/10.1007/S11229-008-9435-2

Igarashi, Y., Nagata, K., Kuwatani, T., Omori, T., Nakanishi-Ohno, Y., & Okada, M. (2016). Three Levels Of Data-Driven Science. *Journal Of Physics: Conference Series*, 699(1), 1-14.

Iwasaki, Y., Sawada, R., Stanev, V., Ishida, M., Kirihara, A., Omori, Y., Someya, H., Takeuchi, I., Saitoh, E., & Yorozu, S. (2019). Identification Of Advanced Spin-Driven Thermoelectric Materials Via Interpretable Machine Learning. *Npj Computational Materials*, 5(1), 1–6. Https://Doi.Org/10.1038/S41524-019-0241-9

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction To Statistical Learning*. Springer.

Kaminski, A., & Schneider, R. (Publicación Pendiente). *Social, Technical, And Mathematical Opacity: Computer Simulation And The Scientific Work On Purification*. 1-36. Https://Philpapers.Org/Rec/KAMSTA-3

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, In My Hand: Who's The Fairest In The Land? On The Interpretations, Illustrations, And Implications Of Artificial Intelligence. *Business Horizons*, 62(1), 15-25.

Kapoor, S., & Narayanan, A. (2022). Leakage And The Reproducibility Crisis In ML-Based Science. *Arxiv:2207.07048*. Https://Doi.Org/10.48550/Arxiv.2207.07048

Kapteyn, M. G., & Willcox, K. E. (2020). From Physics-Based Models To Predictive Digital Twins Via Interpretable Machine Learning. *Arxiv:2004.11356 [Cs]*. Http://Arxiv.Org/Abs/2004.11356

Khalifa, K. (2017). *Understanding, Explanation, And Scientific Knowledge.* Cambridge University Press. Https://Doi.Org/10.1017/9781108164276.

Kiefer, C. (2016). Assessing The Quality Of Unstructured Data: An Initial Overview. En *Proceedings Of The LWDA 2016 Proceedings (LWDA)*, Vol. 1670, 62-73. Http://Ceur-Ws.Org/Vol-1670/Paper-25.Pdf

Kitchin, R. (2014). Big Data, New Epistemologies And Paradigm Shifts. *Big Data & Society*, 1(1), 1-12.

Kitchin, R. (2014b). *The Data Revolution: Big Data, Open Data, Data Infrastructures And Their Consequences*. Sage.

Koh, P. W., & Liang, P. (2017). Understanding Black-Box Predictions Via Influence Functions. *Arxiv Preprint Arxiv:1703.04730*, 1-11.

Koumakis, L. (2020). Deep Learning Models In Genomics; Are We There Yet?. *Computational And Structural Biotechnology Journal*, 18, 1466-1473.

Krause, J., Perer, A., & Ng, K. (2016). Interacting With Predictions: Visual Inspection Of Black-Box Machine Learning Models. *Proceedings Of The 2016 CHI Conference On Human Factors In Computing Systems*, 5686–5697.

Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine Generals Problem. *ACM Transactions On Programming Languages And Systems*, 382–401.

Lange, M. (2013). What Makes A Scientific Explanation Distinctively Mathematical?. *The British Journal For The Philosophy Of Science*, 64, 485-511.

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring Strategies For Training Deep Neural Networks. *Journal Of Machine Learning Research*, 10(1).

Leonelli, S. (2014). What Difference Does Quantity Make? On The Epistemology Of Big Data In Biology. *Big Data & Society*, 1(1), 1-11.

Leonelli, S. (2020). Scientific Research And Big Data. *The Stanford Encyclopedia Of Philosophy*. Edward N. Zalta (Ed.). Https://Plato.Stanford.Edu/Archives/Sum2020/Entries/Science-Big-Data/

Levien, R. E., & Maron, M. E. (1967). A Computer System For Inference Execution And Data Retrieval. *Communications Of The ACM*, 10(11), 715-721.

Lillicrap, T. P., & Kording, K. P. (2019). What Does It Mean To Understand A Neural Network? *Arxiv:1907.06374*.

Lipton, Z. C. (2018). The Mythos Of Model Interpretability. *Queue*, 16(3), 31-57.

Lipworth, W., Mason, P. H., Kerridge, I., & Ioannidis, J. P. (2017). Ethics And Epistemology In Big Data Research. *Journal Of Bioethical Inquiry*, 14(4), 489-500.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, And Emotions*. Cambridge University Press.

Liu, H., & Motoda, H. (Eds.). (2007). *Computational Methods Of Feature Selection*. CRC Press.

Marr, B. (2016). *Big Data In Practice: How 45 Successful Companies Used Big Data Analytics To Deliver Extraordinar*y. John Wiley And Sons Ltd.

Marr, B. (2017, Enero 23). Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud. *Forbes*. Https://Www.Forbes.Com/Sites/Bernardmarr/2017/01/23/Really-Big-Data-At-Walmart-Real-Time-Insights-From-Their-40-Petabyte-Data-Cloud/

Marsh, L. C., & Cormier, D. R. (2001). *Spline Regression Models (No. 137)*. Sage

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is Psychology Suffering From A Replication Crisis? What Does "Failure To Replicate" Really Mean? *American Psychologist*, 70(6), 487–498. Https://Doi.Org/10.1037/A0039400

Mazzocchi, F. (2015). Could Big Data Be The End Of Theory In Science? A Few Remarks On The Epistemology Of Data-Driven Science. *EMBO Reports*, 16(10), 1250-1255.

Mcculloch, W. S., & Pitts, W. (1943). A Logical Calculus Of The Ideas Immanent In Nervous Activity. *The Bulletin Of Mathematical Biophysics*, 5(4), 115-133.

Microsoft. (S/F). *Excel Specifications And Limits*. Microsoft. Https://Support.Microsoft.Com/En-Us/Office/Excel-Specifications-And-Limits-1672b34d-7043-467e-8e27-269d656771c3

Miller, T. (2019). Explanation In Artificial Intelligence: Insights From The Social Sciences. *Artificial Intelligence*, 267, 1–38. Https://Doi.Org/10.1016/J.Artint.2018.07.007

Minsky, M., & Papert, S. A. (2017). *Perceptrons: An Introduction To Computational Geometry*. MIT Press.

Mitchell, M., & Krakauer, D. C. (2023). The Debate Over Understanding In AI's Large Language Models. Proceedings Of The National Academy Of Sciences, 120(13), 1-5. Https://Doi.Org/10.1073/Pnas.2215907120

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations Of Machine Learning*. The MIT Press.

Molina, M., & Garip, F. (2019). Machine Learning For Sociology. *Annual Review Of Sociology*, 45(1), 27–45. Https://Doi.Org/10.1146/Annurev-Soc-073117-041106

Molnar, C. (2022). *Interpretable Machine Learning*. Leanpub. Recuperado El 15 De Febrero De 2022, De Https://Christophm.Github.Io/Interpretable-Ml-Book/

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction To Linear Regression Analysis*. John Wiley & Sons.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations. *Proceedings Of The 2020 Conference On Fairness, Accountability, And Transparency*, 607–617. Https://Doi.Org/10.1145/3351095.3372850

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter Sentiment Classification: The Role Of Human Annotators. *Plos ONE*, 11(5). Https://Doi.Org/10.1371/Journal.Pone.0155036

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, Methods, And Applications In Interpretable Machine Learning. *Proceedings Of The National Academy Of Sciences*, 116(44), 22071–22080.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep Learning Applications And Challenges In Big Data Analytics. *Journal Of Big Data*, 2(1), 1-21.

O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big Data', Hadoop And Cloud Computing In Genomics. *Journal Of Biomedical Informatics*, 46(5), 774-781.

Ostertagová, E. (2012). Modelling Using Polynomial Regression. *Procedia Engineering*, 48, 500-506.

Papernot, N., Mcdaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks Against Machine Learning. *Proceedings Of The 2017 ACM On Asia Conference On Computer And Communications Security*, 506–519. Https://Doi.Org/10.1145/3052973.3053009

Plastino, A. R., Plastino, A., & Tsallis, C. (1994). The Classical N-Body Problem Within A Generalized Statistical Mechanics. *Journal Of Physics A: Mathematical And General*, 27(17), 5707-5714.

Reichenbach, H. (1956). *The Direction Of Time*. University Of California Press.

Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization Of The World From Edge To Core. *IDC*. Https://Www.Seagate.Com/Files/Www-Content/Our-Story/Trends/Files/Idc-Seagate-Dataage-Whitepaper.Pdf

Resch, M., & Kaminski, A. (2019). The Epistemic Importance Of Technology In Computer Simulation And Machine Learning. *Minds And Machines*, 29(1), 9–17. Https://Doi.Org/10.1007/S11023-019-09496-5

Reutlinger, A. (2017). Explanation Beyond Causation? New Directions In The Philosophy Of Scientific Explanation. *Philosophy Compass*, 12(2), 1-11. Https://Doi.Org/10.1111/Phc3.12395

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining The Predictions Of Any Classifier. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, 1135–1144. Https://Doi.Org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings Of The AAAI Conference On Artificial Intelligence, 32(1), Article 1*. Https://Ojs.Aaai.Org/Index.Php/AAAI/Article/View/11491

Rokach, L., & Maimon, O. Z. (2008). *Data Mining With Decision Trees: Theory And Applications*. World Scientific.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain. *Psychological Review*, 65(6), 386.

Ruder, S. (2016). An Overview Of Gradient Descent Optimization Algorithms. *Arxiv:1609.04747*.

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. Https://Doi.Org/10.1038/S42256-019-0048-X

Rudin, C., & Ustun, B. (2018). Optimized Scoring Systems: Toward Trust In Machine Learning For Healthcare And Criminal Justice. *INFORMS Journal On Applied Analytics*, 48(5), 449–466. Https://Doi.Org/10.1287/Inte.2018.0957

Russell, B. (2009). *The Philosophy Of Logical Atomism*. Routledge.

Salmon, W. C. (1984). *Scientific Explanation And The Causal Structure Of The World*. Princeton University Press.

Schmidt, P., & Biessmann, F. (2019). Quantifying Interpretability And Trust In Machine Learning Systems. *Arxiv:1901.08558 [Cs, Stat]*. Http://Arxiv.Org/Abs/1901.08558

Schölkopf, B. (2019). Causality For Machine Learning. *Arxiv:1911.10500 [Cs, Stat]*. Http://Arxiv.Org/Abs/1911.10500

Semenova, L., Rudin, C., & Parr, R. (2022). On The Existence Of Simpler Machine Learning Models. *2022 ACM Conference On Fairness, Accountability, And Transparency*, 1827–1858. Https://Doi.Org/10.1145/3531146.3533232

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable Publications Are Cited More Than Replicable Ones. *Science Advances*, 7(21), Eabd1705. Https://Doi.Org/10.1126/Sciadv.Abd1705

Shrout, P., & Rodgers, J. (2018). Psychology, Science, And Knowledge Construction: Broadening Perspectives From The Replication Crisis. *Annual Review Of Psychology*, 69, 487–510. Https://Doi.Org/10.1146/Annurev-Psych-122216-011845

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A General Reinforcement Learning Algorithm That Masters Chess, Shogi, And Go Through Self-Play. *Science*, 362(6419), 1140–1144. Https://Doi.Org/10.1126/Science.Aar6404

Simsek, Z., Vaara, E., Paruchuri, S., Nadkarni, S., & Shaw, J. D. (2019). New Ways Of Seeing Big Data. *Academy Management Journal*, 62(4), 971-978.

Sivaramakrishnan, R., Antani, S., Candemir, S., Xue, Z., Abuya, J., Kohli, M., Alderson, P., & Thoma, G. (2018). Comparing Deep Learning Models For Population Screening Using Chest Radiography. *Medical Imaging 2018: Computer-Aided Diagnosis*, 10575.

Skopek, J. (2018). Big Big Data's Epistemology And Its Implications For Precision Medicine And Privacy. En *Big Data, Health Law, And Bioethics* (Eds.) Cohen, I. G., Lynch, H. F., Vayena, E., & Gasser, U. Cambridge University Press.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME And SHAP: Adversarial Attacks On Post Hoc Explanation Methods. *Proceedings Of The AAAI/ACM Conference On AI, Ethics, And Society*, 180–186. Https://Doi.Org/10.1145/3375627.3375830

Slack, D., Hilgard, S., Lakkaraju, H., & Singh, S. (2021). Counterfactual Explanations Can Be Manipulated. *Arxiv:2106.02666 [Cs].* Http://Arxiv.Org/Abs/2106.02666

Stathakis, D. (2009). How Many Hidden Layers And Nodes?. *International Journal Of Remote Sensing*, 30(8), 2133–2147.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical Or Genomical? *PLOS Biology*, 13(7), 1-11.

Sullivan, E. (2019). Understanding From Machine Learning Models. *The British Journal For The Philosophy Of Science*, 73(1), 1-34.

Tanaka, A., Tomiya, A., & Hashimoto, K. (2021). *Deep Learning And Physics*. Springer Singapore. Https://Doi.Org/10.1007/978-981-33-6108-9

Tansley, S., & Tolle, K. M. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. A. J. Hey (Ed.). Redmond. *Microsoft Research*.

Trovati, M., Hill, R., Anjum, A., Zhu, S. Y., & Liu, L. (Eds.). (2015). *Big-Data Analytics And Cloud Computing*. Springer International Publishing.

Tukey, J. W. (1962). The Future Of Data Analysis. *The Annals Of Mathematical Statistics*, 33(1), 1-67.

Van Der Aalst, W. (2016). Data Science In Action. En Van Der Aalst (Ed.), *Process Mining: Data Science In Action* (Pp. 3–23). Springer.

Van Dijck, J. (2014). Datafication, Dataism And Dataveillance: Big Data Between Scientific Paradigm And Ideology. *Surveillance & Society*, 12(2), 197-208.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press

Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread Of True And False News Online. *Science*, 359(6380), 1146–1151. Https://Doi.Org/10.1126/Science.Aap9559

Watson, D. S., & Floridi, L. (2020). The Explanation Game: A Formal Framework For Interpretable Machine Learning. *Synthese*. Https://Doi.Org/10.1007/S11229-020-02629-9

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., Mcinnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical Applications Of Machine Learning Algorithms: Beyond The Black Box. *British Medical Journal*, 364, 1886

Wheeler, G. (2017). "Machine Epistemology And Big Data", En Mcintyre & Rosenberg (Eds.), *The Routledge Companion To Philosophy Of Social Science*, Routledge, 321-329.

Wiener, J., & Bronson, N. (2014). Facebook's Top Open Data Problems. *Facebook Research*. Https://Research.Fb.Com/Blog/2014/10/Facebook-S-Top-Open-Data-Problems/

Wiggins, B. J., & Christopherson, C. D. (2019). The Replication Crisis In Psychology: An Overview For Theoretical And Philosophical Psychology. *Journal Of Theoretical And Philosophical Psychology*, 39(4), 202–217. Https://Doi.Org/10.1037/Teo0000137

Williamson, T. (2017). Model-Building In Philosophy. En R. Blackford & D. Broderick (Eds.), *Philosophy's Future* (1a Ed., Pp. 159–171). Wiley. Https://Doi.Org/10.1002/9781119210115.Ch12

Wittgenstein, L. (2012). *Tractatus Logico-Philosophicus*. Alianza Editorial

Woodward, J. & Ross, L. (2021). *Scientific Explanation*. The Stanford Encyclopedia Of Philosophy. N. Zalta (Ed.). Https://Plato.Stanford.Edu/Archives/Sum2021/Entries/Scientific-Explanation/

Woodward, J. (2003). *Making Things Happen*. Oxford University Press.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How Do Visual Explanations Foster End Users' Appropriate Trust In Machine Learning? *Proceedings Of The 25th International Conference On Intelligent User Interfaces*, 189–201. Https://Doi.Org/10.1145/3377325.3377480

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding The Effect Of Accuracy On Trust In Machine Learning Models. *Proceedings Of The 2019 CHI Conference On Human Factors In Computing Systems*, 1–12. Https://Doi.Org/10.1145/3290605.3300509

Zednik, C., & Boelsen, H. (2022). Scientific Exploration And Explainable Artificial Intelligence. *Minds And Machines*, 32(1), 219–239. Https://Doi.Org/10.1007/S11023-021-09583-6

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency In Algorithmic And Human Decision-Making: Is There A Double Standard? *Philosophy & Technology*, 32(4), 661–683. Https://Doi.Org/10.1007/S13347-018-0330-6

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why Should You Trust My Explanation?" Understanding Uncertainty In LIME Explanations. *Arxiv:1904.12991 [Cs, Stat]*. Http://Arxiv.Org/Abs/1904.12991

ÍNDICE ANALÍTICO

A

B

C

Ch

C

D

E

F

G

H

I

T

V

W

Y

Z