APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO AL ANÁLISIS ECONÓMICO Y LA FORMULACIÓN DE POLÍTICAS

Stephen HANSEN
Universidad de Oxford

Resumen

Este artículo explora las diversas aplicaciones de los algoritmos de aprendizaje automático (o *machine learning*) al análisis económico y la formulación de políticas. A tal fin, comienza introduciendo los conceptos básicos del aprendizaje automático y distinguiendo las dos grandes ramas en las que se desdobla: el aprendizaje supervisado y el no supervisado. En la segunda parte del artículo se analizan ejemplos de utilización y aplicaciones prácticas de los algoritmos de aprendizaje automático. Una de dichas aplicaciones es la cuantificación de datos no estructurados y la recuperación de dicha información de un modo que resulte útil para los economistas. La segunda de las aplicaciones analizadas se refiere a las nuevas posibilidades de medición, donde la combinación del aprendizaje automático y los nuevos datos digitales ofrece la oportunidad de desarrollar parámetros referidos a variables como la inflación y la actividad económica. Las últimas dos aplicaciones tienen que ver con la formulación de proyecciones (*forecasting*) y con la inferencia causal. El mensaje general del artículo es que el aprendizaje automático ofrece las herramientas necesarias para sacar partido de las enormes posibilidades de las nuevas fuentes de datos digitales.

Palabras clave: aprendizaje automático, digitalización de datos.

Abstract

This article focuses on applications of machine learning algorithms for economic research and policymaking. It first introduces basic concepts in machine learning, whose main branches include supervised and unsupervised learning. The second half of the article discusses use cases and applications of machine learning algorithms. First, it discusses the quantification of unstructured data and how to recover information in a way that is useful for economists. The second application concerns new possibilities for measurement, where the combination of machine learning and new digital data, provides the opportunity to develop measures of objects like inflation and economic activity. The last two applications are related to forecasting and causal inference. The overall message of the article is that machine learning provides the tools needed to fully exploit the possibilities of rich new digital data sources.

Key words: machine learning, digital data.

JEL classification: C55.

I. INTRODUCCIÓN

En los últimos años hemos observado un crecimiento espectacular de la generación de datos. El volumen de datos generado en 2014 y 2015 superó al acumulado hasta ese momento en épocas anteriores, y se estima que para 2020 habrá unos 44 *zettabytes,* o sea, 44 billones de *gigabytes,* de datos (Marr, 2015). Gran parte de esta explosión se debe a la digitalización, pues las nuevas tecnologías permiten almacenar actividades humanas que anteriormente tenían naturaleza efímera. Hoy es habitual enviar mensajes y fotografías en formato electrónico o subirlos a las redes sociales, lo que a su vez permite almacenarlos en servidores de manera indefinida. Datos digitales de una relevancia económica más directa están ahora cada vez más disponibles. Por ejemplo, información relacionada con las compras de los consumidores individuales, historiales detallados sobre los precios de los productos o minuciosos registros administrativos están empezando a transformar el análisis empírico en economía.

Esta avalancha de datos ha venido acompañada de nuevos métodos empíricos para analizarlos. Como respuesta a la digitalización de los datos, el aprendizaje automático (o *machine learning*) se ha desarrollado a pasos agigantados en los últimos diez años, y ha nutrido de muchas ideas a la inteligencia artificial, campo que en la actualidad suscita gran interés entre el público. Menos clara es la relevancia de todos estos avances para la investigación empírica en economía. La mayoría de los métodos de aprendizaje automático los han desarrollado científicos computacionales, estadísticos e ingenieros, cuyos objetivos suelen diferir de los objetivos de los economistas a la hora de formular trabajos empíricos. Esto plantea la cuestión de la utilidad potencial del aprendizaje automático en el ámbito de la economía, dado el énfasis que ésta última pone en la inferencia causal y en la predicción contrafactual.

Este artículo tiene un doble objetivo: primero, ofrecer una introducción a los conceptos básicos del aprendizaje automático, lo que realiza en su parte inicial. Y segundo, reflexionar sobre el impacto potencial del aprendizaje automático en la investigación económica y en las políticas públicas, y hacerlo a través de una discusión, no tanto de carácter técnico, sino centrada en nociones generales (1), de sus diversas áreas de aplicación.

Se llega a varias conclusiones. En primer lugar, una aplicación importante del aprendizaje automático, pero que no suele resaltarse, es su capacidad para usar tipos de datos completamente nuevos. La econometría moderna utiliza normalmente

datos «regulares», entendiendo por tales aquellos representables en forma de tabla, con las observaciones individuales en las filas y las variables en columnas. Además, las variables suelen registrarse como mediciones cuantitativas únicas: por ejemplo, el gasto total de las familias o los salarios de los trabajadores. Sin embargo, muchas de las nuevas fuentes de datos digitales disponibles no tienen este formato; el texto, las imágenes de satélite o los perfiles de búsquedas en Internet contienen amplias cantidades de información relevante económicamente, pero presentan una estructura de datos no estándar. El aprendizaje automático puede utilizarse para extraer la información importante de estas fuentes y filtrarla de cara al análisis econométrico. El artículo ilustra varios casos en los que se han utilizado diferentes enfoques estándar o de serie para hacer esto de manera efectiva.

En segundo lugar, es importante reconocer que, a menudo, los métodos de aprendizaje automático *no* resultan adecuados para las clases de problemas a los que se enfrentan los economistas. El artículo cita ejemplos de esto en lo que respecta a la elaboración de proyecciones y a la inferencia causal.

En tercer lugar, pese a la diferencia de objetivos del aprendizaje automático y la economía, ideas concretas del *machine learning* pueden adoptarse, y extenderse, para satisfacer las necesidades del análisis económico. Este proceso se encuentra aún en sus estadios iniciales en la economía, pero es probable que tenga la llave para permitir a los economistas y responsables de políticas explotar plenamente el potencial de los datos digitales.

II. ¿QUÉ ES EL APRENDIZAJE AUTOMÁTICO?

No parece existir una definición unánimemente aceptada de aprendizaje automático. A modo de introducción, podemos definirlo como el estudio de algoritmos que permiten a las máquinas mejorar su rendimiento en alguna tarea determinada conforme se le van suministrando nuevos datos. Para ser más precisos, un conocido manual lo expresa como «un conjunto de métodos diseñados para detectar automáticamente patrones en los datos, y luego utilizar dichos patrones detectados para predecir datos futuros, o para tomar otras clases de decisiones bajo condiciones de incertidumbre» (Murphy, 2012). Ahora bien, ninguna de estas definiciones acaba de transmitir del todo las diferencias entre el aprendizaje automático y la econometría. A fin de cuentas, el modelo de regresión de mínimos cuadrados ordinarios, conocido de sobra por cualquier estudiante de economía, detecta relaciones en los datos y genera estimaciones de mayor calidad al aplicarse sobre bases de datos amplias.

Un punto que diferencia al aprendizaje automático de la econometría es el papel que juega la inferencia estadística. La econometría tiende a centrarse en procedimientos formales de inferencia. Esto conlleva estimar los parámetros de un determinado modelo estadístico, y derivar las propiedades teóricas de estas distribuciones para testar hipótesis. El aprendizaje automático, en cambio, suele preocuparse menos del «verdadero» modelo que genera los datos y buscar, en su lugar, procedimientos que simplemente funcionen bien bajo alguna métrica, como la precisión predictiva. Esta distinción no es tajante. Por ejemplo, algunos algoritmos de aprendizaje automático (en particular, bayesianos) parten, a semejanza de la econometría, de un modelo de probabilidad asumido para los datos, y dicho modelo puede en principio utilizarse con fines de inferencia. No obstante, incluso en estos casos, la literatura sobre el aprendizaje automático suele preocuparse menos de las garantías de inferencia teórica que la literatura sobre econometría. Breiman (2001) aporta una buena introducción a estas «dos culturas» de modelización estadística.

Otra área de diferenciación es la computación. Los procedimientos econométricos rara vez se evalúan en términos de su complejidad computacional, mientras que tales consideraciones forman el zócalo de gran parte del aprendizaje automático. La popularidad de determinados algoritmos estriba precisamente en que son rápidos de calcular y son escalables. Esto se debe básicamente a la enorme amplitud de las bases de datos utilizadas en muchas aplicaciones de aprendizaje automático. Los economistas pueden permitirse trabajar con algoritmos computacionalmente ineficientes dado el tamaño mucho más reducido de las bases de datos que suelen manejar, pero esto evolucionará a medida que dichas bases de datos crezcan en dimensión.

También existen algunas diferencias semánticas que, en ocasiones, oscurecen lo que en realidad son ideas similares. Ambos campos formulan modelos que tratan de explicar ciertas variables de interés, denominadas algebraicamente *y*, mediante algunas

otras variables potencialmente relacionadas con *y*, a las que se denomina *x*. En econometría, es habitual referirse a *y* como la «variable dependiente», o el «objeto estudiado», y designar a las *x* con el nombre de «regresores», o variables «explicativas» o «independientes». En el aprendizaje automático, la *y* recibe a menudo el nombre de «etiqueta» o «respuesta», mientras que para referirse a la *x* se emplean términos como «características» o «predictores». Además, al proceso de construcción de un modelo que explique *y* a partir de *x* se le denomina en econometría «estimación», mientras que en aprendizaje automático se llama «aprendizaje». En este artículo se seguirá la terminología estándar de la econometría.

Pero en lugar de debatir sobre la definición exacta del aprendizaje automático, es preferible plantearse las tareas específicas que se pretende resolver con él. Una clasificación típica es la que distingue entre *aprendizaje supervisado* y *aprendizaje no supervisado*, cuyo significado explicamos a continuación.

1. Aprendizaje supervisado

El aprendizaje supervisado consiste en construir un modelo para explicar una variable de estudio a partir de ciertas variables explicativas. Esto es exactamente lo que hacen muchos modelos econométricos, pero la métrica para juzgar la calidad de un modelo en aprendizaje automático es bastante distinta. Esencialmente, el único objetivo es la precisión predictiva. Lograr una alta precisión predictiva con una base de datos fija es una tarea simple. Un modelo de regresión lineal que para explicar la variable dependiente usase tantas variables explicativas como observaciones existen en la muestra, conseguiría un ajuste perfecto a los datos. Procedimientos como estos, sin embargo, tienden a sobreajustar y hacer predicciones basadas en relaciones espurias. El aprendizaje automático aspira, por tanto, a la precisión predictiva de las observaciones extramuestrales (*out-of-sample*). El objetivo es construir un modelo que sea capaz de predecir acertadamente el valor de la variable explicada en los nuevos datos que no fueron utilizados para la construcción del modelo. Los modelos que consiguen buenos resultados en esto se consideran eficaces.

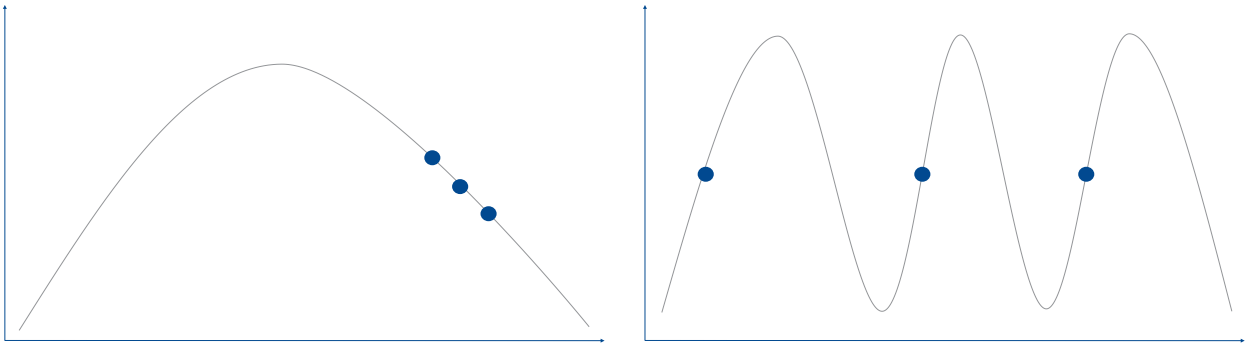
Podemos verlo con un ejemplo concreto. Supongamos el caso del correo basura (coloquialmente, *spam*). La variable dependiente que se desea estudiar es binaria: o bien un *e-mail* es *spam* o no lo es. Las variables independientes son las palabras incluidas en los correos electrónicos. Dado un conjunto finito de correos electrónicos, predecir el correo basura es potencialmente tan simple como encontrar una palabra que solo se encuentre presente en los correos *spam* y nunca en los que no lo son. Supongamos que esa palabra es «xxx». Entonces, la presencia de «xxx» es un predictor perfecto de correo basura en este conjunto específico de *e-mails.* Pero este modelo no puede generalizarse bien a los nuevos *e-mails,* por ejemplo, a correos basura en los que se solicite a una persona los datos de su cuenta bancaria para ingresarle una supuesta herencia si atiende a su petición. En su lugar, deseamos un modelo que sea capaz de distinguir los *nuevos* correos y clasificarlos correctamente como *spam* o no *spam*.

La literatura sobre aprendizaje automático ha avanzado a pasos agigantados en la construcción de modelos con buena capacidad predictiva. Los algoritmos de reconocimiento facial en las fotos, de reconocimiento de voz y el anteriormente citado problema de detección del correo basura se utilizan actualmente de manera generalizada en la sociedad, y son todos ellos aplicaciones del aprendizaje supervisado.

Aun cuando se desease utilizar la precisión predictiva como referente para juzgar la eficacia de un modelo, existen dudas respecto a si el modo de evaluar los algoritmos de aprendizaje supervisado es suficiente. ¿Cómo se puede evaluar el desempeño de un algoritmo respecto a datos extramuestrales si tales datos no están disponibles? La solución estándar consiste en dividir los datos en dos subgrupos: una muestra de entrenamiento y una muestra de prueba. La muestra de entrenamiento se utiliza para estimar el modelo. Luego, por cada observación en los datos de prueba, se puede generar un valor de predicción para el resultado dado el modelo estimado con la muestra de entrenamiento, y a continuación comparar la predicción contra el valor real en los datos de prueba. La muestra de prueba representa los datos extramuestrales, ya que no fue utilizada en el entrenamiento. Sin embargo, a menudo no hay garantías de que los datos manejados en el grupo de prueba coincidan con los datos extramuestrales *reales* con los que se verá confrontado el algoritmo en el mundo real. El gráfico 1 que figura a continuación aporta una ilustración. Supongamos la situación de la izquierda. Los datos observados son los tres puntos sobre la curva, y estamos tratando de predecir un resultado medido sobre el eje vertical dada cierta variable explicativa,

GRÁFICO 1

PELIGRO DE DATOS NO REPRESENTATIVOS PARA LA EVALUACIÓN DEL APRENDIZAJE SUPERVISADO



representada en el eje horizontal. La curva traza la relación existente en el mundo real entre la variable explicativa y la que se trata de explicar. Un algoritmo de aprendizaje supervisado construido solamente con los tres puntos observados podría tener un muy mal ajuste *incluso si* consigue una elevada precisión predictiva extramuestral con un grupo de prueba. Esto se debe a que todos los datos provienen de una parte restringida de la curva que se comporta como una línea con pendiente descendente, y un algoritmo supervisado tenderá a estimar sencillamente ese patrón. Dicho patrón, obviamente, no ajusta bien cuando se generaliza a todos los valores posibles de la variable explicativa, puesto que parte de la relación existente en el mundo real tiene una pendiente creciente. De forma similar, en la situación de la derecha, los datos observados darán nuevamente una imagen engañosa de la verdadera relación. El problema es en este caso que los datos observados están demasiado dispersos (2).

Estos ejemplos son sencillos e implican una variable explicativa unidimensional. En las aplicaciones reales del aprendizaje automático, existen centenares, miles, o incluso millones de *inputs* diferentes, y determinar si los datos sobre los que se evalúa a los algoritmos supervisados ofrecen una visión representativa del mundo es extremadamente complicado. Los economistas y responsables de políticas deberían tener esto en cuenta. Mientras que los errores en el reconocimiento de imágenes o de voz pueden ser molestos y embarazosos, tienen escasos costes sociales. Los errores en el diseño y la formulación de políticas pueden tener consecuencias catastróficas.

Además, por lo general, los algoritmos de aprendizaje supervisado están construidos en entornos bastante diferentes a los que se encuentran los economistas. En primer lugar, son ricos en datos. Empresas como Facebook y Google pueden recurrir a bases ingentes de datos para entrenar sus algoritmos de recomendación. En cambio, los economistas a menudo tienen entre manos un conjunto muy limitado de datos con los que trabajar. Por ejemplo, aunque predecir una recesión es un problema de política importante, las recesiones se presentan con escasa frecuencia en las series temporales históricas. En segundo lugar, los entornos son estables, en el sentido de que el futuro se parece mucho al pasado. Pero las economías no suelen ser estacionarias, y lo son aún menos cuando la precisión predictiva reviste mayor importancia, por ejemplo, al inicio de una crisis financiera o en la introducción de una tecnología disruptiva. Esto pone en entredicho si los métodos de aprendizaje automático estándar son adecuados para el tipo de problemas de predicción en los que están interesados los economistas. El artículo volverá sobre esta cuestión más adelante durante la discusión de las aplicaciones.

Para entender mejor las diferencias entre el aprendizaje automático y la econometría tradicional, es instructivo considerar un popular algoritmo supervisado llamado *LASSO* (*Least Absolute Shrinkage and Selection Operator*), introducido por Tibshirani (1996) y que ha visto crecer su aceptación en economía (véase, por ejemplo, Belloni *et al*. 2014). El *LASSO* es una extensión básica del modelo de regresión por mínimos cuadrados ordinarios (MCO) que constituye una de las

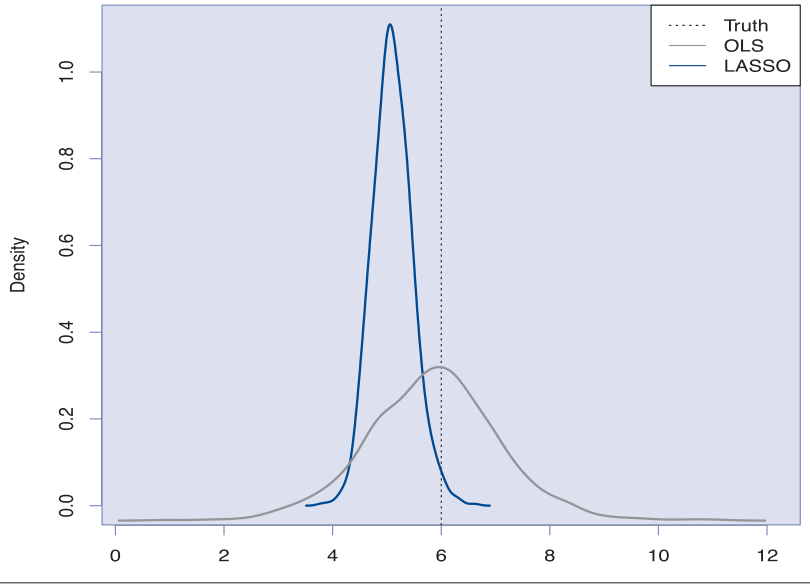
piedras angulares de la economía aplicada. Ambos modelos relacionan una variable dependiente (*y)* con variables independientes (*x)* eligiendo coeficientes para los valores de *x* que mejor explican *y*. Por ejemplo, *y* podría ser el nivel de renta, y *x* podría estar formado por tres variables: número de años de formación académica, cociente intelectual y color de ojos. Cabe esperar que las dos primeras variables estén relacionadas con la renta, no así la tercera. La diferencia clave entre el método MCO y el *LASSO* es que el *LASSO* añade un término que penaliza los valores de coeficiente alto. La lógica subyacente es asignar un coeficiente cero a las variables poco importantes y un coeficiente distinto de cero a las variables importantes. Al hacerlo, se espera que las variables con coeficientes distintos de cero tengan relación con el fenómeno que se pretende explicar, y que las que reciben coeficientes de cero sean variables sin relevancia alguna, es decir, *ruido*. En el ejemplo anterior, esto significaría que el *LASSO* asignaría un coeficiente positivo a los años de educación y al cociente intelectual, y un coeficiente igual a cero al color de ojos. Dicho enfoque puede ser particularmente fructífero cuando el número de variables existentes es alto en relación con el número de observaciones. De hecho, es posible estimar el *LASSO* incluso cuando hay muchas más variables que observaciones.

Aunque el término de penalización en el *LASSO* puede eliminar las variables que generan *ruido*, ello tiene un coste. El término de penalización «castiga» los valores de coeficiente alto para *todas* las variables independientes. Esto significa que los coeficientes, incluso de las variables significativas, son más bajos de lo que serían en un modelo simple de MCO. En la jerga de la econometría, los valores de los coeficientes estimados en *LASSO* adolecen, por tanto, de sesgo: el efecto estimado de cualquier variable independiente es, de media, de menor magnitud que su efecto real. Siguiendo con el ejemplo anterior, supongamos que un año adicional de educación genera una renta adicional de 600 euros al año. El *LASSO* podría estimar que el efecto adicional de un año de educación es solo de 300 euros al año. Siendo esto así, ¿por qué querríamos utilizar un modelo que deliberadamente introduce un sesgo en su procedimiento de estimación? La respuesta es que introducir un sesgo reduce el ruido. El modelo MCO estimará cierto coeficiente para el color de ojos, incluso si esta variable no está relacionada en absoluto con la renta. De media, el valor del coeficiente estimado será cercano a cero, pero, dependiendo de la aleatoriedad que exista en una base de datos concreta, el MCO podría reflejar cierta correlación espuria entre el color de ojos y la renta. A su vez, esto introduce *ruido* en la predicción que haga el modelo sobre la renta. En cambio, el *LASSO* simplemente tenderá a dejar fuera del modelo el color de ojos.

El gráfico 2 ilustra estas propiedades. Supongamos una persona con un historial académico de cinco años y un cociente intelectual de 100. Además, supongamos que un año adicional de educación incrementa la renta en 0,6 unidades, y que un punto adicional de cociente intelectual incrementa la renta en 0,03 unidades. Así, el ingreso real de este individuo es igual a 5 * 0,6 + 100 * 0,03 = 6. El gráfico 2 muestra un esquema de la distribución de los valores de la variable dependiente resultante de la predicción de MCO y *LASSO* cuando en el modelo existen muchas variables irrelevantes (o *ruido)* sin relación alguna con la

GRÁFICO 2

COMPROMISO ENTRE SESGO Y DISPERSIÓN



renta. Aquí vemos claramente lo que en la literatura sobre el aprendizaje automático se conoce como el compromiso entre sesgo y dispersión (*bias-variance tradeoff*). De media, el MCO genera la predicción correcta, ya que la distribución tiene su punto central en 6. Pero, alrededor de dicho punto, observamos una gran dispersión, con valores de predicción tan bajos como 0 hacia un extremo y tan altos como 12 hacia el otro. El *LASSO,* en cambio, está sesgado, pues el valor central de la distribución es 5, en lugar de 6. Pero las predicciones muestran una mayor concentración alrededor del valor central de 5, sin llegar a los valores extremos del MCO. Dicho de otro modo, el *LASSO* falla en la predicción media, pero nunca se equivoca demasiado; el MCO, en cambio, acierta en promedio, pero con frecuencia se equivoca mucho. Se puede demostrar en este ejemplo que el error cuadrático medio —una métrica popular para medir la bondad del ajuste— es inferior en el *LASSO* que en el MCO.

¿Qué implicaciones se derivan de este ejemplo? Numerosos manuales de econometría limitan su atención a aquellos modelos que son, de media, correctos (sin sesgo), buscando dentro de dichos modelos los que presentan menor dispersión. El aprendizaje automático nos enseña que este enfoque podría ser restrictivo, en especial cuando existen muchas variables y cuando el objetivo principal es la predicción, en cuyo caso los modelos con sesgo pueden comportarse bien. Al mismo tiempo, como se ha dicho, los economistas están interesados en modelos con buena capacidad para la inferencia: al decidir la cantidad para invertir en escuelas públicas, es crucial conocer el efecto real que un año adicional de educación tiene sobre la renta (0,6 en el ejemplo anterior). Puesto que los algoritmos de aprendizaje supervisado están diseñados para la precisión predictiva, surge lógicamente la pregunta de si existe una relación inversa entre los dos objetivos. Dicho de otro modo, ¿pueden utilizarse los algoritmos de aprendizaje supervisado para la inferencia de parámetros, incluso si no fueron diseñados con ese objetivo en mente? En muchos casos importantes, la respuesta es «no» o, quizá sea más exacto decir, «no a menos que se introduzca alguna modificación». Como hemos visto en el caso del *LASSO*, las estimaciones de los coeficientes tienen un sesgo a la baja. Además, no hay garantía de que el *LASSO* omita todas las variables irrelevantes (*ruido*). Existe cierto corpus teórico sobre inferencia estadística con el *LASSO* (los lectores interesados pueden consultar Bühlmann y van de Geer, 2011; o Hastie, Tibshirani y Wainwright, 2015), pero en la práctica hay pocas garantías fiables que sean consistentes de unas aplicaciones a otras.

El principal mensaje que podemos extraer es que el aprendizaje supervisado ha logrado recientemente grandes avances en lo que respecta a la predicción extramuestral en entornos estables y ricos en datos. Con frecuencia, lo ha conseguido introduciendo cierto sesgo para reducir la dispersión, lo que es crucial en modelos con un amplio número de variables. Pero en qué medida y en qué casos dichos modelos pueden utilizarse para los problemas de inferencia que interesan a muchos economistas, sigue siendo una cuestión sin resolver y materia de estudio activa. En una sección posterior sobre aplicaciones discutiremos las recientes contribuciones.

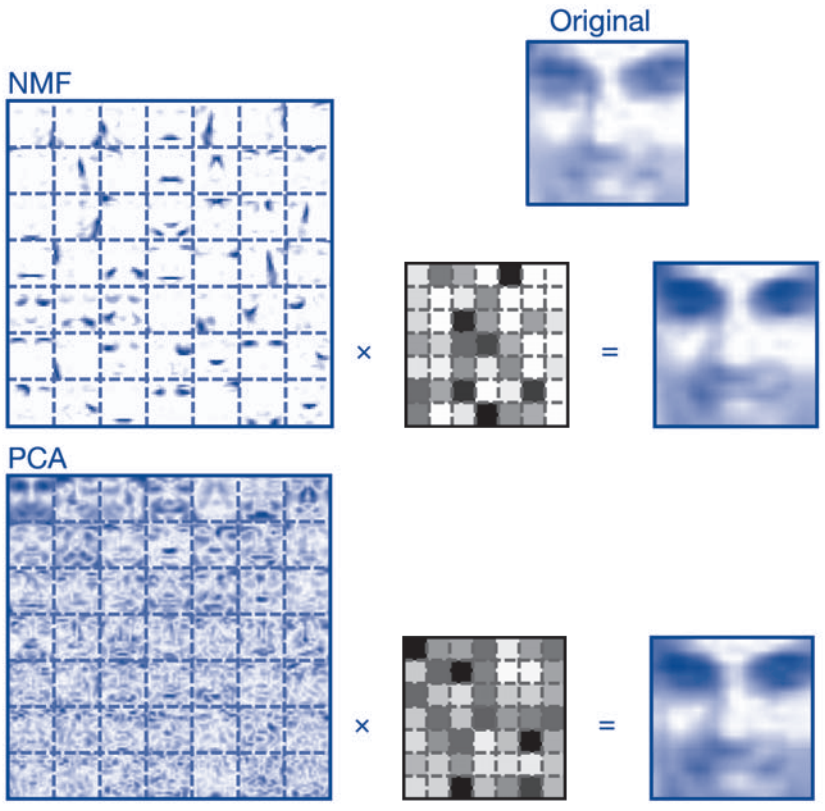
2. Aprendizaje no supervisado

Aunque el aprendizaje no supervisado ha recibido algo menos de atención en la literatura, es un campo importante por derecho propio. El objetivo del aprendizaje no supervisado es poner de manifiesto la estructura oculta en los datos. En este caso, no hay una variable dependiente que se pretenda explicar mediante variables independientes. Cada observación en un conjunto de datos simplemente está integrada por múltiples variables registradas, con interdependencias potencialmente complejas, que el aprendizaje no supervisado trata de desvelar. Pueden existir varias motivaciones para el aprendizaje no supervisado: una es describir las fuentes de mayor capacidad explicativa dentro de un *pool* amplio de variables independientes. Alternativamente, el aprendizaje no supervisado puede proporcionar una representación con dimensionalidad reducida de un objeto con alta dimensionalidad, preservando, al mismo tiempo, la mayoría de la información relevante. El aprendizaje no supervisado también puede agrupar las observaciones en función de sus similitudes. Ninguna de estas motivaciones debería resultar totalmente desconocida para los economistas. El *clustering* y el análisis factorial, por ejemplo, constituyen tareas de aprendizaje no supervisado que ya se practican bastante en la economía empírica.

El aprendizaje no supervisado puede ser un fin en sí mismo cuando la exploración de los datos es el objetivo primario; alternativamente, puede verse como un medio para preparar los datos utilizados para extraer características que sirvan de *inputs* a los algoritmos de aprendizaje supervisado o a los modelos econométricos. En análisis

GRÁFICO 3

ANÁLISIS DE LOS COMPONENTES PRINCIPALES FRENTE A LA FACTORIZACIÓN DE MATRIZ NO NEGATIVA



Fuente: extraído de Lee y Seung (1999).

económico aplicado, ello lo hace más fácilmente digerible que el aprendizaje supervisado. Para bien o para mal, los aspectos de inferencia formal son a menudo descartados, incluso en economía, cuando el objetivo primario es el procesamiento y la preparación de datos. En este sentido, los métodos estándar de aprendizaje no supervisado pueden aplicarse más fácilmente si aportan una descripción de los datos más completa que los métodos existentes. Durante el resto de la discusión tratamos de demostrar que éste es el caso.

Probablemente, el algoritmo de aprendizaje no supervisado más conocido en economía es el análisis principal de componentes (*PCA*, por sus siglas en inglés). La idea es encontrar factores o componentes comunes entre variables que expliquen cómo se mueven conjuntamente. Las observaciones son, así pues, representadas como combinaciones de estos factores comunes, en lugar de en términos de las variables originales. Los investigadores suelen utilizar un número de factores mucho menor que variables hay para representar las observaciones, de manera que existe una reducción de la dimensionalidad. Por ejemplo, tal enfoque es frecuente en las series temporales macroeconómicas para explicar el comovimiento de cientos de índices económicos diferentes. Los factores comunes pueden verse como variables cíclicas no observadas que determinan los datos observados.

El *PCA* es también bastante conocido en el ámbito de la literatura sobre aprendizaje automático, si bien éste ha desarrollado algoritmos adicionales que corrigen algunas de las limitaciones del *PCA*. Aunque los economistas no suelen ser conscientes de ellos, pueden incorporarse a la caja de herramientas econométricas a un coste bastante bajo. Una limitación del *PCA* es que los factores que identifica pueden ser difíciles de interpretar, y, en muchos casos, parecerse más a objetos abstractos que explican comovimientos en lugar de a objetos con sentido real. En la literatura sobre aprendizaje automático, hay trabajos que exploran formas alternativas de construir factores que resuelvan este problema en algunas aplicaciones. Un ejemplo interesante se encuentra en Lee y Seung (1999), quienes comparan el *PCA* con una alternativa a la que denominan factorización de matriz no negativa (*NMF*, por sus siglas en inglés). Este último método es similar al *PCA*, salvo por la restricción que impone de que los factores solo puedan ser cifras no negativas. Esta distinción, aparentemente técnica, es en realidad sustantiva, pues los factores que genera el *NMF* se parecen más a las partes elementales de las que procede cada observación de los datos.

El gráfico 3 ilustra esta idea para datos de imágenes. La base de datos subyacente consiste en una colección de fotografías

de rostros humanos. Las matrices más completas de siete por siete, a la izquierda de la figura, ilustran los 49 componentes que el *PCA* y el *NMF* revelan de las fotografías (el sombreado azul oscuro indica cifras positivas, y el sombreado gris indica cifras negativas). Lo fascinante del ejemplo es que los componentes del *NMF* se parecen a elementos del rostro: hay ojos, bocas, narices, etc. A continuación se construye una fotografía única de la muestra combinando estos elementos para dar lugar a un rostro individual (las matrices más pequeñas en el centro de las figuras, muestran las ponderaciones específicas a cada fotografía aplicadas a los componentes para llegar a la observación de la derecha). Los componentes del *PCA* son muy diferentes: el primer componente es básicamente un rostro promedio, y el resto de los componentes añaden y sustraen intensidad de píxeles a este rostro promedio. Entonces se representa un rostro específico como una desviación ponderada frente al rostro promedio, lo que supone una construcción menos intuitiva de la que ofrece el *NMF*.

Este ejemplo puede parecer una curiosidad, pero ilustra una cuestión más esencial de la que los economistas podrían beneficiarse para conocer la estructura latente a partir del uso de los algoritmos comunes en aprendizaje automático, hasta la fecha casi completamente ignorados. Por ejemplo, el *NMF* y los algoritmos relacionados se podrían aplicar a las ventas de productos individuales por grupos de consumidores, para conocer patrones de compra arquetípicos e identificar bienes sustitutivos y complementarios; o a los precios de productos individuales, con el fin de determinar los factores subyacentes de la inflación.

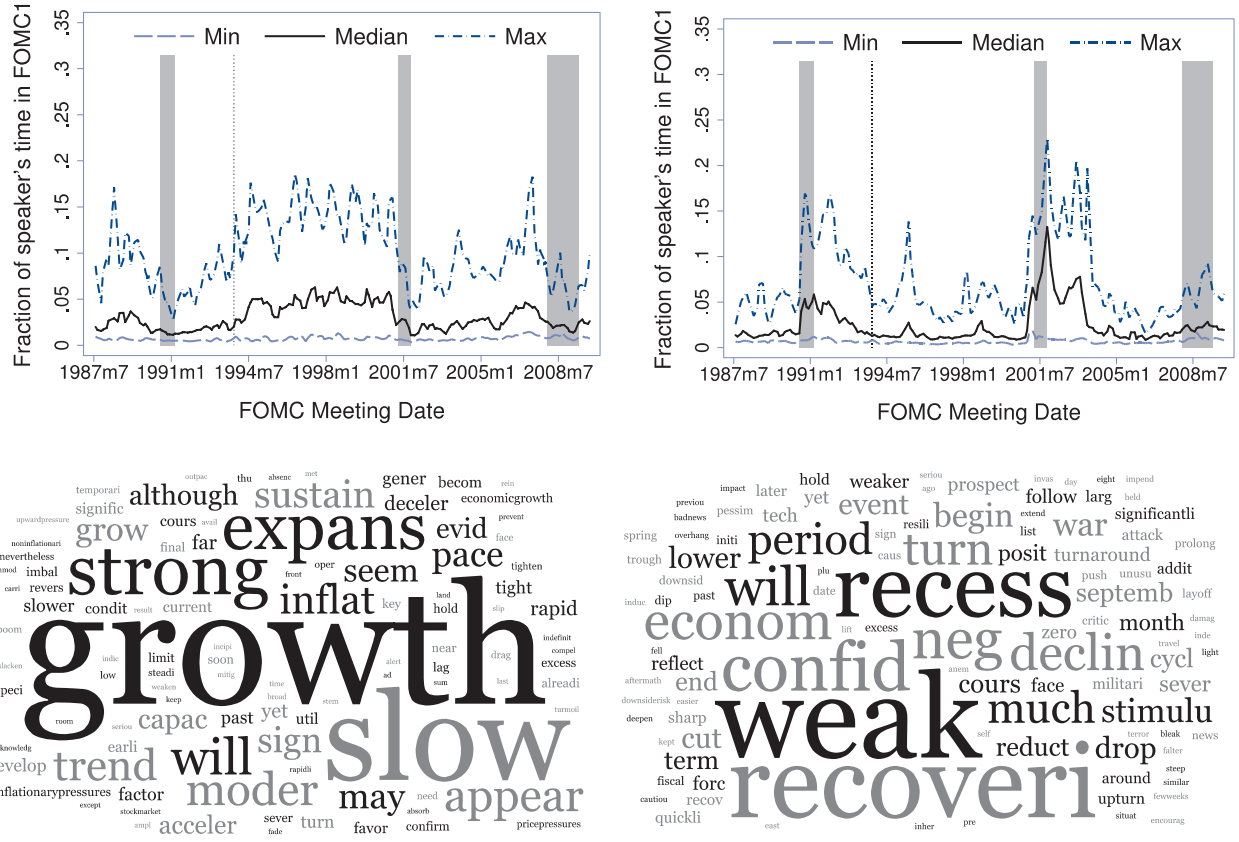
Otra limitación del *PCA* es que sus fundamentos se adaptan mejor a los datos que varían de manera continua. Un ejemplo importante de datos para los que esto no se cumple es el texto. La forma más básica de representar bases de datos textuales, también llamadas «corpora», es contar la ocurrencia de todos los términos únicos del vocabulario en todos los documentos. Los datos resultantes tienen claramente interdependencias, por ejemplo, la palabra «trabajo» tenderá a presentarse junto con la palabra «salario». Pero los datos son fundamentalmente discretos: una palabra no puede aparecer 1,5 veces. Además, hay una gran mayoría de palabras únicas en los «corpora» que no aparecen en ningún documento específico, de modo que los datos también están poblados en un gran porcentaje de ceros. Dichos datos requieren algoritmos que modelicen sus características concretas.

Uno de los modelos de aprendizaje no supervisado más potentes y populares para texto es el *Latent Dirichlet Allocation* (*LDA*), introducido por Blei, Ng y Jordan (2003). El *LDA* es un ejemplo de un modelo temático probabilístico, que identifica temas en los «corpora» y luego representa los documentos como combinaciones de esos temas. Más concretamente, un tema es una distribución de probabilidad asociada a todas las palabras únicas en el corpus. Este aspecto probabilístico del *LDA* es importante. Supongamos que alguien imagina un tema sobre inflación y otro sobre desempleo. Tomemos ahora la palabra «tasa». A primera vista, no está claro a cuál de los temas debería pertenecer «tasa», pues ese término podría aparecer en un tema sobre inflación o en otro sobre desempleo, con las expresiones «tasa de inflación» y «tasa de participación en el mercado laboral», respectivamente. Permitir la asociación probabilística de palabras a temas transmite esta flexibilidad semántica. El *LDA* es también un modelo de pertenencia mixta, porque los documentos no se asocian a un único tema. En su lugar, a cada documento se le asignan porcentajes de todos los temas. Así, un documento puede ser un 25 por 100 sobre desempleo, un 10 por 100 sobre inflación, etcétera.

El gráfico 4 muestra un ejemplo de *output de LDA* estimado a partir de un corpus de transcripciones literales de las reuniones del Comité Federal de Mercado Abierto (*FOMC*, por sus siglas en inglés), el órgano responsable de decidir la política monetaria en Estados Unidos. El período muestral para la estimación es 1987-2009. Las dos nubes de palabras representan dos temas diferentes objeto de estimación. El tamaño de cada palabra en la nube refleja aproximadamente la probabilidad de su ocurrencia en el tema (3). Aunque el algoritmo no se alimenta con ninguna información sobre el contenido subyacente de los datos, los temas son claramente interpretables: uno versa sobre crecimiento económico, el otro sobre recesión y recuperación. La serie temporal encima de cada tema muestra la variación en la proporción de tiempo que los miembros individuales del *FOMC* dedican a discutir los respectivos temas (la línea discontinua azul es la proporción máxima de tiempo en una reunión determinada; la línea continua negra es la proporción de tiempo mediana, y la línea discontinua azul claro es la proporción mínima). Los períodos de recesión están sombreados en color gris. La serie también muestra propiedades

GRÁFICO 4

EJEMPLO DE *OUTPUT* DEL *LATENT DIRICHLET ALLOCATION*



Fuente: Hansen, McMahon y Prat, 2018.

muy naturales. La atención al crecimiento aumenta sistemáticamente cuando la economía está en fase expansiva, y se desploma al comienzo de los períodos de recesión. En contraste, la atención dedicada a la recesión repunta durante las contracciones de la actividad económica. Nuevamente, merece la pena señalar que tales patrones han sido capturados completamente por un algoritmo de aprendizaje automático, sin contribución alguna del investigador.

Otro punto importante es la alta dimensionalidad innata de los textos. Los «corpora», incluso los de tamaño moderado, contienen miles de términos únicos. El sobreajuste de dichos datos es un problema grave, pero la estructura estadística del *LDA* protege frente a esto. Se trata de lo que se conoce como un modelo *bayesiano*, lo que significa que otorga cierta probabilidad inicial a todas las posibles combinaciones de palabras en cada tema. Los datos observados cambian dichas probabilidades, pero sin llegar a determinarlas por completo. La citada base de datos de transcripciones posee aproximadamente 10.000 términos únicos, y aun así el *LDA* maneja la dimensionalidad sin apuros.

Estos dos ejemplos muestran el poder del aprendizaje no supervisado para captar patrones interesantes implícitos en los datos. Además, muestra cómo el aprendizaje automático puede convertir lo que a primera vista es una maraña de datos no estructurados —es decir, archivos de imagen y texto sin depurar— en formas cuantitativas manejables y adecuadas para el análisis estadístico tradicional. Esto abre la puerta no solo a la posibilidad de nuevas técnicas que aplicar a los datos

existentes, sino al acceso a nuevos datos. Esta cuestión se desarrolla más ampliamente en la sección sobre aplicaciones que aparece a continuación.

Una posible crítica a los algoritmos de aprendizaje automático es que adolecen de falta de estructura. El gráfico 3 sugiere que las probabilidades de la cobertura de los temas podrían no ser estáticas en el tiempo, sino depender de la fase del ciclo económico, pero dicho aspecto no está integrado en el *LDA*. Una posible contribución de los economistas al desarrollo de algoritmos de aprendizaje automático no supervisado es introducir en ellos dependencias de interés que vinculen más directamente sus *outputs* a los atributos de interés. Tales esfuerzos requerirán probablemente de colaboración entre disciplinas.

III. APLICACIONES

Explicadas las bases conceptuales del aprendizaje automático, en el resto del artículo se abordan sus aplicaciones potenciales al campo de la economía y a la formulación de políticas. Comenzamos con una de las aplicaciones más pragmáticas: la cuantificación de datos nuevos y seguidamente en formas manejables. A continuación, consideramos el papel que el aprendizaje automático juega en la conversión de datos digitales en medidas económicas más específicas, y seguidamente una discusión del aprendizaje automático en los modelos de elaboración de proyecciones *(forecasting). Finalmente, reflexionamos sobre posibles aplicaciones para la inferencia causal.*

1. Cuantificación de datos no estructurados

Muchas empresas y reguladores poseen cantidades masivas de datos no estructurados, en particular datos de texto. Un ejemplo es el sector de la abogacía, en el que una buena parte del trabajo de los letrados que se inician en la práctica jurídica consiste en rastrear infinidad de documentos en busca de contenido relevante, desde contratos y escrituras notariales a jurisprudencia anterior, etc. Los reguladores afrontan una tarea similar cuando inician expedientes supervisores. Por ejemplo, las redadas policiales para perseguir posibles delitos contra la competencia suelen aportar montañas de documentos, y separar el material relevante del irrelevante supone una ardua tarea. Automatizar la tarea de encontrar información relevante tiene, por tanto, el potencial de generar grandes ganancias de eficiencia en estos contextos, y dicho proceso se encuentra ya bastante avanzado en el sector judicial (Croft, 2017).

Una de las formas más comunes de determinar la relevancia del documento en economía es la búsqueda por palabras clave. Según este enfoque, se empieza definiendo una palabra, o una lista de ellas, y a continuación los documentos son señalizados como que contienen dicho términos o no, o bien se ordenan de acuerdo con la frecuencia con la que aparecen. Pese a su relativa sencillez y fácil implementación, las búsquedas por palabras clave tienen sus limitaciones. La más obvia es que requieren de una definición previa de las palabras importantes, lo que puede exigir juicios subjetivos. Por ejemplo, para medir la actividad económica, podríamos elaborar una lista de palabras, entre ellas «crecimiento». Pero no cabe duda de que existen otras palabras que también se emplean para discutir la actividad, y elegirlas implica numerosos juicios subjetivos. Un argumento más sutil es que el «crecimiento» también aparece mencionado en otros contextos, como al describir el crecimiento de los salarios como foco de presiones inflacionistas, y en la práctica es muy difícil tener en cuenta el contexto al realizar búsquedas por palabras clave. En otros casos, el académico o el responsable de políticas simplemente podría no tener ni idea de cómo se relacionan las palabras con el contenido de su interés. En los litigios relacionados con la manipulación de precios de mercado por los operadores, como en el reciente escándalo por el acuerdo para fijar el valor del tipo interbancario de oferta de Londres (*LIBOR*, por sus siglas inglés), gran parte de la evidencia procede de charlas entre operadores en las que estos utilizan jerga, argot y lenguaje codificado que convierte la búsqueda simple por palabras clave en una tarea difícil de implementar.

El aprendizaje automático no supervisado ayuda a superar algunos de estos problemas. Especialmente cuando hay incertidumbre sobre el contenido de los documentos, y sobre el lenguaje utilizado en contextos concretos, el aprendizaje automático aporta un enfoque potente impulsado por los datos para la exploración de corpus y la recuperación de información. La cuantificación de datos no estructurados podría ser un fin en sí mismo, por ejemplo, permitiendo a un regulador cribar rápidamente los documentos y clasificarlos en categorías. O podría ser la primera fase en la extracción de características de los datos textuales que utilizar luego como *inputs* para ulteriores estudios empíricos.

A fin de ilustrar mejor estos puntos, supongamos una obser-

GRÁFICO 4

OBSERVACIÓN INCLUIDA EN LAS TRANSCRIPCIONES DEL FOMC

We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Fuente: Janet Yellen, marzo 2006.

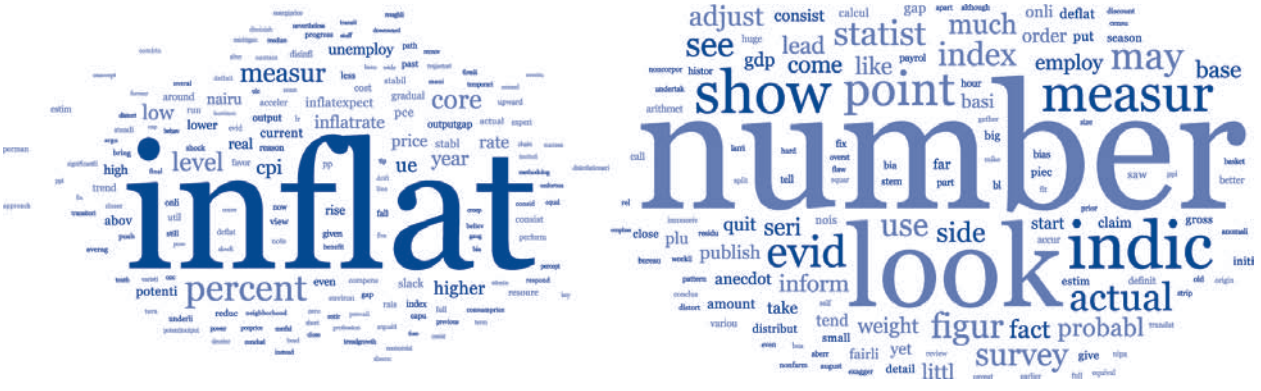
vación, incluida en el corpus de transcripciones del *FOMC* antes mencionado, que se reproduce a continuación en el gráfico 4. Se trata de un párrafo de Janet Yellen pronunciado en marzo de 2006 cuando era presidenta del Banco de la Reserva Federal de San Francisco. Este párrafo incluye un lenguaje muy técnico, y determinar su contenido manualmente requeriría que el lector contase con un alto nivel de conocimientos de economía.

Como alternativa al procesamiento manual, se puede utilizar el *LDA*, el algoritmo de aprendizaje no supervisado descrito antes, para determinar su contenido. El modelo *LDA* estimado asocia esta manifestación sobre todo al tema de la izquierda en el gráfico 5 más abajo. Este tema otorga a su vez la máxima probabilidad a «inflación» (y otras palabras que comienzan con la raíz «inflat»). Lo fascinante de esta ilustración es que, en la observación del ejemplo, no hay ninguna cita literal de la palabra «inflación», y una búsqueda por palabras clave no la habría identificado como relevante. En cambio, Janet Yellen utiliza muchas palabras relacionadas con inflación (*CPI* es el índice de precios al consumo; *PCE* es el deflactor del gasto personal), y el *LDA* aprende, a partir de *otros* documentos que forman parte del corpus, que las palabras que Yellen utiliza se inscriben en la mayoría de los casos en situaciones en las que *también* se emplea la palabra «inflación». Esto permite asociar esta manifestación al tema inflación.

Otro punto de interés es que el *LDA* es capaz de contextualizar adecuadamente las palabras individuales dentro de los documentos. Tomemos la palabra *measures* («medidas») utilizada por Yellen en el ejemplo. Aunque esta palabra aparece de manera destacada en el tema inflación, también está presente, con una alta probabilidad, en otro tema sobre indicadores numéricos que se muestra a la derecha en el gráfico 5. El *LDA* consigue resolver esta ambigüedad prestando atención a otras palabras pronunciadas por Yellen. Aunque la clasificación de «medidas» fuera de contexto no es inmediata, el

GRÁFICO 5

TEMA MÁS ASOCIADO CON EL EJEMPLO (IZDA.) Y TEMA ALTERNATIVO (DCHA.)



hecho de que Yellen utilice muchas palabras inequívocamente asociadas con el tema inflación le lleva a asignar en este caso «medidas» al tema inflación.

Estas características del *LDA* ayudan a explicar su gran popularidad. Una aplicación familiar para muchos lectores podría ser el sistema de indexación en JSTOR, el conocido repositorio de artículos académicos. Gracias al *LDA*, los lectores reciben alertas de nuevos artículos potencialmente interesantes en el repositorio tomando como referencia el contenido estimado en los artículos consultados previamente por ellos. La adopción de tales sistemas por empresas y reguladores que también manejan amplios «corpora» textuales podría reportar grandes beneficios en la recuperación automática de información.

Aunque el foco de la discusión ha girado hasta ahora en torno al texto, sus conclusiones también son válidas para otras clases de datos no estructurados. Los organismos de políticas y las empresas de *marketing* recopilan periódicamente datos de encuestas para medir actitudes, comportamientos y características. Estos datos suelen analizarse de forma *ad hoc*, por ejemplo, computando la respuesta promedio a una batería de preguntas para llegar a un único valor numérico. Nuevamente, el aprendizaje no supervisado proporciona un medio de modelizar la estructura de dependencias completa en los datos y de extraer nuevas perspectivas sobre las diferencias subyacentes entre los encuestados. Un ejemplo en la literatura económica es Bandiera *et al*. (2017), que analiza encuestas detalladas sobre el uso del tiempo a más de 1.000 consejeros delegados de una variedad de países. Utilizando el *LDA*, establecen una novedosa distinción conductual entre los *CEO* «líderes», que dedican tiempo a coordinar las funciones de alto nivel de la empresa, y los «gestores», que dedican tiempo a cuestiones más operativas del negocio. Enfoques similares se han utilizado para medir el estatus de salud (Erosheva, Fienberg y Joutard, 2007) y la ideología política (Gross y Manrique-Vallier, 2014) a partir de datos de encuestas.

Otra intrigante aplicación potencial del aprendizaje no supervisado consiste en los datos de red, donde el reto es identificar grupos de nodos relacionados basándose en patrones de vinculación. Existe una nutrida literatura sobre este «problema de detección de comunidad», como así se denomina, fuera de la economía, pero apenas hay aplicaciones a la economía. Una excepción es Nimczik (2017), que, mediante aprendizaje no supervisado, estima el alcance geográfico de los mercados laborales utilizando datos de flujos de trabajadores en Austria.

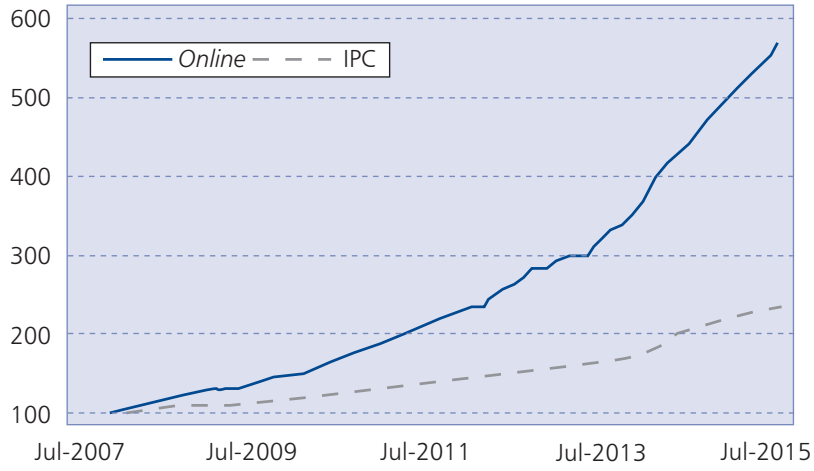
2. Nuevos datos y nueva medición

La primera aplicación referida en esta sección consistió simplemente en utilizar el aprendizaje automático para dar sentido a un conjunto de datos deslavazado y difícil de interpretar imponiendo una estructura mínima en el proceso de recuperación de información. Sin embargo, existe un creciente interés no solo en describir tales datos, sino también en utilizarlos para construir nuevas medidas de variables económicas relevantes. Los indicadores económicos tradicionales están limitados en varios aspectos: por ejemplo, a menudo están disponibles a intervalos temporales relativamente distantes, como sucede con las mediciones del producto interior bruto, de periodicidad trimestral. Además, suelen estar construidos para unidades geográficas agregadas, como Estados nación, con poca granularidad espacial. Por último, en muchas regiones del mundo, las estadísticas económicas oficiales o bien no están disponibles en su totalidad, o bien están manipulada por los gobiernos hasta el punto de contener muy escasa información. Por estos motivos, hay una demanda de nuevas fuentes de información. Recientemente, se ha constatado un creciente interés en los datos digitales como medio de cubrir estas lagunas. Entre otros ejemplos, cabe citar:

– En Argentina, el gobierno manipuló activamente las estadísticas oficiales de precios a comienzos de 2007. El proyecto *The Billion Prices Project* desarrollado en el Massachusetts Institute of Technology se diseñó como una manera de ofrecer un índice de inflación alternativo más preciso utilizando precios publicados por los comerciantes *online* en Argentina, y se ha expandido a muchos otros países desde entonces. Aunque el universo de distribuidores minoristas para los que se pueden obtener precios *online* es menor que el observado por las agencias gubernamentales oficiales, estos precios se actualizan a diario, tienen un bajo coste de extracción y están libres de interferencias gubernamentales. A continuación, el gráfico 6 contrasta las medidas de inflación calculadas a partir de precios *online* y las estadísticas oficiales de precios, y demuestra la capacidad de los datos digitales para capturar la dinámica subyacente real en una economía cuando no se dispone

GRÁFICO 6

INFLACIÓN ARGENTINA MEDIDA POR PRECIOS ONLINE Y LECTURA OFICIAL DEL IPC



Fuente: Cavallo y Rigobon, 2016.

de datos oficiales o estos son poco fiables.

– Baker, Bloom y Davis (2016) construyen un popular e influyente índice denominado *Economic Policy Uncertainty* (EPU) (http://www.policyuncertainty.com/). Aunque la importancia de la incertidumbre para la actividad económica es indiscutible, históricamente ha habido pocos indicadores adecuados de la incertidumbre. Parámetros basados en los mercados financieros, como el VIX, se calculan utilizando los precios de las opciones obtenidos de los mercados de renta variable estadounidenses, los cuales no capturan toda la incertidumbre que afrontan los agentes económicos. En cambio, el índice *EPU* mide específicamente la incertidumbre que rodea la formulación de políticas. En una gran parte, se construye a partir de la fracción de artículos publicados en una amplia selección de periódicos que contienen términos como «incierto», «económica», «congreso» y «regulación».

– Glaeser, Kim y Luca (2017) construyen un índice de actividad local utilizando el número de restaurantes y negocios reseñados en el sitio web Yelp. Este índice tiene capacidad predictiva para los datos, mucho más agregados y rezagados, del US Census Bureau sobre patrones de actividad económica por condados, referidos en particular a las zonas más densamente pobladas.

– SpaceKnow es una empresa comercial que genera numerosos índices de actividad económica utilizando datos de imágenes por satélite. Uno de dichos índices es el *China Satellite Manufacturing Index,* que se basa en 2.200 millones de instantáneas individuales de más de 6.000 polígonos industriales en China (Wigglesworth, 2018).

Aunque los índices de actividad son algunos de los objetos de interés más obvios que pueden facilitar los nuevos datos, también hay otros que los son menos, pero brindan posibilidades igualmente potentes. Un ejemplo es el trabajo de Hoberg y Phillips (2010, 2016) y tiene relevancia directa con la política sobre competencia. La cuestión abordada es cómo medir la clasificación industrial de las empresas. Los sistemas de clasificación SIC o NAICS, ampliamente utilizados, tienen varias limitaciones. Las empresas no suelen recibir diferentes clasificaciones a lo largo del tiempo, aun si los mercados en los que operan cambian. Los sistemas de clasificación son no demasiado buenos a la hora de replicar la evolución de productos completamente nuevos. Y, con carácter más general, proporcionan una visión poco refinada de cómo se diferencian las empresas entre sí.

Hoberg y Phillips proponen utilizar datos de texto para construir clasificaciones sectoriales que solventen algunos de estos retos. La idea es utilizar las descripciones de productos contenidas en las declaraciones anuales corporativas (impreso 10-K) ante la Securities and Exchange Commission estadounidense. Por cada par de empresas que realizan una declaración en cada año, puede computarse una medida de similitud lingüística entre sus respectivas descripciones, y utilizarla como sustituto de la proximidad entre ambas empresas en el espacio de productos. Además, a partir de estas medidas de similitud, es posible agrupar las empresas en *clusters* para definir categorías sectoriales. La clasificación resultante aporta una medida dinámica y continua del lugar que las empresas ocupan en el espacio de

productos respecto a todas las demás empresas de la muestra. Hoberg y Phillips muestran que su clasificación basada en el texto brinda varias perspectivas novedosas sobre los motivos que llevan a las empresas a fusionarse y sobre cómo se desarrollan nuevos productos.

En este punto, resulta conveniente hacer una distinción entre el *big data* procedente de fuentes digitales, por un lado, y el aprendizaje automático, por el otro. Si bien los datos digitales «en bruto» o sin procesar contienen, sin duda, información relevante para las variables económicas de interés, es difícil establecer una correspondencia exacta (mapeo) entre ambas esferas. Una posibilidad es aplicar algoritmos de aprendizaje no supervisado para describir los datos según lo analizado en la primera aplicación, y luego utilizar las características extraídas para construir un índice de interés. El problema es que estas características no habrán sido elegidas por su máxima capacidad predictiva de la variable económica que se desea explicar, lo que implica una pérdida de información y, por tanto, de eficacia.

En cambio, la tarea de construir nuevos índices a partir de bases de datos amplias supone, en muchos aspectos, un problema clásico de aprendizaje supervisado, puesto que el objetivo primario es conseguir la mejor predicción posible de la variable objeto de estudio. Jean *et al*. (2016) es un ejemplo de investigación que combina extensas bases de datos digitales (imágenes por satélite) y algoritmos de aprendizaje automático supervisado de vanguardia para obtener una nueva medición económica (niveles de pobreza, granulares desde el punto de vista espacial, en varios países de África). A medida que se extienda el uso del aprendizaje automático en economía, muchos de los índices construidos a partir de datos digitales constituirán probablemente el *output* de algoritmos supervisados perfilados.

3. Elaboración de proyecciones (*forecasting*)

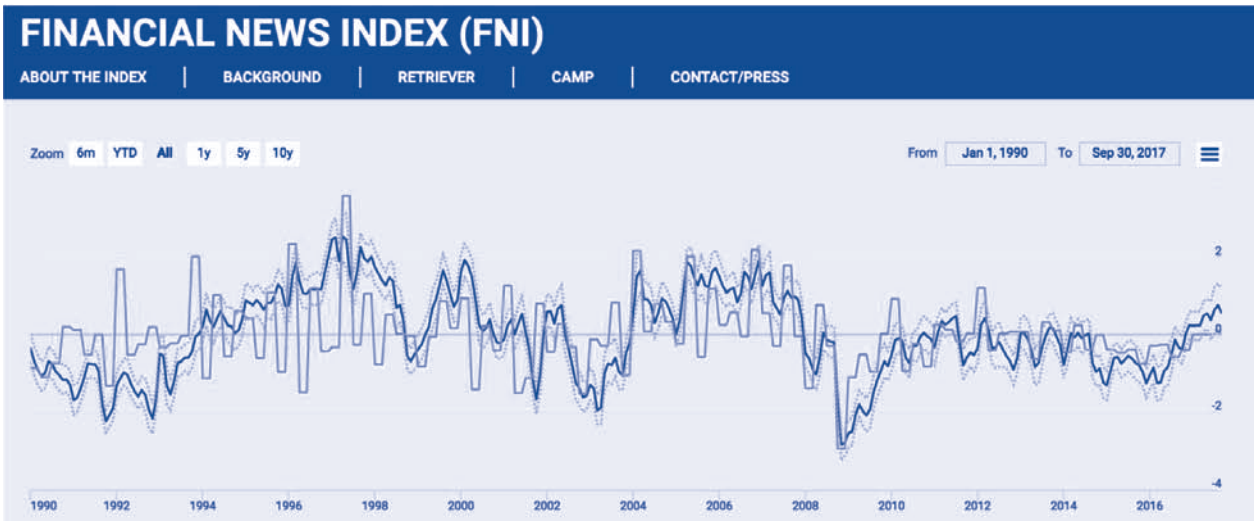
Según se ha explicado anteriormente, el aprendizaje automático supervisado consiste, en el fondo, en el estudio de métodos para lograr buenas predicciones extramuestrales a partir de datos de alta dimensionalidad o no estructurados. Un área de gran interés para los responsables de políticas es la elaboración de proyecciones, es decir, la predicción del futuro basándose en datos del pasado. De hecho, la posibilidad de utilizar series temporales de datos económicos para obtener mejores proyecciones del futuro está en el origen del renovado interés en el aprendizaje automático. Stock y Watson (1999) y Bernanke, Boivin y Eliasz (2005) son dos contribuciones seminales a este campo de la literatura que demuestran que el refuerzo de los modelos clásicos de proyección macroeconómica con series temporales amplias puede mejorar los resultados del ejercicio de proyección. Los artículos citados utilizan métodos como la regresión penalizada y la reducción de dimensionalidad, dos técnicas típicas del aprendizaje automático.

Antes de que los economistas apliquen algoritmos de aprendizaje supervisado más modernos a las proyecciones, conviene resaltar de nuevo que la elaboración de proyecciones económicas difiere, en aspectos fundamentales, del entorno en que se construyen y evalúan muchos algoritmos de aprendizaje automático. Primero, una hipótesis habitual en aprendizaje automático es que los datos extramuestrales siguen la misma distribución estadística que los datos de entrenamiento del modelo. Esto, aplicado a un contexto de series temporales, equivale a la hipótesis de que «el futuro se parece al pasado». Si bien en ocasiones puede ser así, hay otras en que podría no serlo si existen cambios estructurales fundamentales. Por ejemplo, cuando tiene lugar un desplazamiento en la capacidad productiva de la economía, la relación histórica entre el desempleo y el crecimiento salarial también cambiará. Aunque existe una consolidada literatura sobre econometría que trata la detección de discontinuidades estructurales, la literatura sobre aprendizaje automático en esta área está mucho menos desarrollada. Segundo, en economía, los datos suelen ser grandes en algunas dimensiones pero pequeños en otras. Aunque existen cientos de series temporales disponibles para elaborar proyecciones, la frecuencia de observación de muchas de ellas es solo trimestral, o tienen una periodicidad incluso menor. Tercero la denominada ratio *signal-to-noise* en los datos económicos y financieros puede ser bastante baja, lo que significa que las relaciones fundamentales entre las variables pueden ser difíciles de detectar porque existe un alto componente de aleatoriedad que afecta a todas las variables del modelo. El gran reto, por tanto, es encontrar maneras de utilizar métodos de aprendizaje supervisado en situaciones para las que no fueron diseñados originalmente.

Una posibilidad es utilizar los denominados modelos *genera-*

GRÁFICO 7

ÍNDICE DE NOTICIAS FINANCIERAS (NEGRO) Y CRECIMIENTO DEL PIB NORUEGO (AZUL)



Fuente: https://www.retriever-info.com/fni

tivos. Estos modelos especifican un modelo estadístico completo para los datos de entrada y de salida, en contraste con algunas tendencias recientes en aprendizaje automático, como el *deep learning*, que adopta una postura más agnóstica sobre el modelo generador de los datos. La principal razón del éxito de los modelos de *deep learning* es su remarcable capacidad predictiva en presencia de amplios datos. Sin embargo, en muestras más reducidas, como las que manejan los economistas, los modelos generativos han demostrado tener mejor capacidad predictiva (Ng y Jordan, 2002). Davig y Smalter Hall (2017) utilizan este hallazgo, y demuestran que un modelo *generativo* consigue predecir mejor las recesiones en EE.UU. que los modelos estándar de regresión y que la *Survey of Professional Forecasters.* Otra ventaja de los modelos generativos es que están más cerca de los modelos estructurales que ya utilizan los economistas con fines de interpretación y estimación.

Otro posible enfoque para formular proyecciones con grandes bases de datos es utilizar primero aprendizaje no supervisado para extraer características, y luego emplear dichas características como *inputs* en un modelo de proyección económica estándar en todo lo demás. Un ejemplo es Thorsrud (2016), que aplica el *LDA* sobre artículos de medios de comunicación noruegos, y utiliza los temas extraídos para predecir la evolución del ciclo económico. El gráfico 7 compara el índice obtenido con el producto interior bruto noruego real. Claramente, el movimiento de las dos series presenta una sustancial correlación, lo que ilustra el valor de las características extraídas a partir del aprendizaje no supervisado con fines de elaboración de proyecciones.

Otro ejemplo ajeno a la macroeconomía es la predicción de conflictos, que es importante tanto a los efectos de la gestión del riesgo por las empresas del sector privado como por los gobiernos. Mueller y Rauh (2017) muestran que los datos de los medios de comunicación pueden ayudar a pronosticar los brotes de violencia política. También utilizan el *LDA* para extraer temas de textos, y luego muestran que la variación en cómo usan los temas de cobertura los periódicos de los países predice los conflictos en dichos países.

Un comentario general aplicable al método de extracción de características y su utilización como *inputs* para los modelos de proyecciones es que, implícitamente, los tratan como datos

fijos en lugar de como objetos estimados. Aunque esto ha conducido a importantes avances en investigación, en el futuro cabría esperar que se desarrollen algoritmos que modelicen conjuntamente datos de alta dimensionalidad y la variable que se está tratando de predecir. Es probable que esto se traduzca en mejores predicciones, y también en una inferencia estadística más rigurosa. De nuevo, los modelos generativos pueden conformar la espina dorsal de dichos planteamientos.

4. Inferencia causal

Las aplicaciones analizadas representan pasos importantes en el trabajo empírico en economía, pero actualmente la profesión está dominada por el interés en la inferencia causal, y más exactamente, en determinar el efecto de las intervenciones por parte de las autoridades de políticas. La utilidad de los modelos predictivos para este fin no es inmediatamente obvia. Athey (2017) presenta una interesante ilustración de este punto. Supongamos que a una cadena de hoteles le interesa determinar el efecto en sus ventas de una subida del precio de la habitación. Si tomáramos simplemente los precios observados y los datos de habitaciones ocupadas, se observaría una relación positiva, porque cuando las tasas de ocupación aumentan los hoteles elevan el precio de las habitaciones libres: durante los picos vacacionales, hay escasez de habitaciones y los precios son altos, mientras que en temporada baja ocurre lo contrario. Por tanto, un modelo puramente predictivo indicaría unas mayores ventas tras un incremento imprevisto de los precios. Por supuesto, el sentido común dice que la relación es justamente la inversa, es decir, un hotel tendría menos ocupación si decidiera subir de improviso los precios de la habitación. El problema aquí es que un modelo predictivo puro basado en datos observados no es capaz de identificar la demanda subyacente no observada de alquiler de habitaciones de hotel. Altas tasas de ocupación están asociadas a precios elevados porque hay una demanda elevada que impulsa ambos a la vez. Los métodos para resolver problemas como estos han sido materia de amplio estudio por parte de la econometría moderna.

Entonces, ¿qué puede ofrecer el aprendizaje automático a los economistas interesados en estimar relaciones causales? Una cuestión importante a tener en cuenta es que incluso los procedimientos de inferencia causal recurren a lo que son esencialmente mecanismos de predicción puros. Un enfoque clásico de inferencia causal es el uso de las denominadas variables «instrumentales». Estas son variables que están correlacionadas con un tratamiento, pero no con el objeto que se desea estudiar (4). Sustituir los instrumentos por el tratamiento permite aislar el impacto causal del tratamiento en el objeto estudiado. La estimación de variables instrumentales suele seguir un proceso en dos fases: primero, se predice el valor del tratamiento dados los instrumentos; segundo, el valor del tratamiento anticipado por el modelo se utiliza como variable independiente en una regresión sobre el objeto a explicar. El primer paso en este procedimiento puede verse como una tarea natural de aprendizaje automático, por cuanto que implica realizar una predicción óptima del tratamiento dados los instrumentos. Los métodos de aprendizaje automático para las variables instrumentales son especialmente relevantes cuando existen muchos instrumentos potenciales, o cuando se desea estimar una relación flexible entre instrumentos y tratamientos. Varios trabajos recientes combinan métodos de aprendizaje automático supervisado con variables instrumentales (Belloni *et al*., 2012; Hartford *et al*., 2017).

Otra aplicación del aprendizaje automático a la inferencia causal es el problema de los controles de alta dimensionalidad. Muchas variables *observables* potenciales también pueden afectar a la variable estudiada más allá del tratamiento elegido. Por ejemplo, el impacto de la cualificación de los trabajadores en la productividad podría depender de las características personales de los trabajadores, de la empresa y de la tecnología que opera el trabajador. No suele estar claro qué variables de control adicionales al tratamiento incluir en la regresión, sobre todo en ausencia de una teoría relevante. Un enfoque común consiste en «correr» muchos modelos diferentes, cada uno de los cuales incluye controles diferentes, y examinar cómo de sensible es la relación entre un tratamiento y la variable estudiada a la inclusión de un grupo particular de controles. Un enfoque de aprendizaje automático poco sofisticado sería incluir todos los controles junto con el tratamiento en un modelo de regresión penalizado para que los datos revelen qué controles son relevantes. De hecho, este enfoque arroja estimaciones poco fiables del efecto del tratamiento, pero los ajustes mediante algoritmos estándar

pueden corregir ese problema (Belloni *et al*., 2014).

Otro enfoque en inferencia causal en economía es la denominada modelización estructural, en la que se toma un modelo económico teórico y luego se utilizan datos para estimar los parámetros de la teoría. A medida que los modelos crecen en complejidad, el número de parámetros puede aumentar con rapidez. Por ejemplo, un modelo sobre demanda del consumidor podría en teoría utilizar elasticidades-precio cruzadas por cada par de bienes posible en un supermercado. El aprendizaje automático también puede ofrecer técnicas de estimación paramétrica en modelos estructurales a gran escala especificados sobre bases de datos a gran escala. Los modelos generativos de formulación bayesiana nuevamente proporcionan un marco natural para la estimación estructural en economía. Aun reconociendo que en los últimos años estos han perdido adeptos entre la comunidad de aprendizaje automático en favor del *deep learning*, su futuro en economía es prometedor. Un ejemplo reciente es Athey *et al*. (2018), si bien esta aplicación del aprendizaje automático es probablemente la menos desarrollada de cuantas se han analizado aquí.

Como en la aplicación a la formulación de proyecciones, vuelve a plantearse la cuestión genérica de que el contexto en el que se fundamentan los algoritmos de aprendizaje automático no tiene necesariamente aplicación directa al campo empírico. Esto no quiere decir que el aprendizaje automático carezca de cualquier relevancia para la inferencia causal, sino que, en esta área, se requiere una evaluación especialmente cuidadosa de en qué casos el aprendizaje automático pueden añadir valor.

IV. CONCLUSIONES

El presente artículo ha revisado los conceptos básicos del aprendizaje automático y ha aportado numerosos ejemplos de cómo puede resultar útil para los economistas académicos y los responsables de formulación de políticas. Algunas aplicaciones simplemente requieren métodos estándar, mientras que otras requieren el desarrollo de nuevas técnicas para abordar los retos específicos de la economía. Aunque algunas de estas técnicas ya están siendo desarrolladas, queda aún mucho trabajo por hacer.

Si bien el eje principal del artículo ha sido el valor que puede derivarse para los responsables de políticas de aplicar técnicas de aprendizaje automático a los datos, también hay nuevas cuestiones regulatorias que han surgido a consecuencia del mayor uso del aprendizaje automático. Un ejemplo de estas cuestiones es la utilización por las empresas de algoritmos de fijación de precios. Cuando las empresas establecen sus precios en función de las características y la conducta de los consumidores individuales, la discriminación de precios, casi necesariamente, aumenta. No está claro si esto reduce el excedente del consumidor. Por un lado, elevar los precios manteniendo constante la cantidad reduce el excedente, pero, por otro, los algoritmos de precios podrían permitir a las empresas aumentar la cantidad o el surtido de bienes producidos. Una segunda cuestión es si el uso de algoritmos de precios puede aumentar la colusión tácita al proporcionar nuevas oportunidades a las empresas de vincular sus precios a los publicados por sus competidores. Esta cuestión ha despertado el interés reciente tanto en el ámbito académico (Salcedo, 2015) como de los responsables de políticas (OCDE, 2017). Pese a la creciente concienciación sobre estas cuestiones, determinar las respuestas apropiadas de las autoridades de la competencia sigue siendo un tema a debate, aunque hay coincidencia en que «el auge de los algoritmos de precios y el *software* de inteligencia artificial exigirá que modifiquemos nuestras prácticas de supervisión del cumplimiento de las leyes» (McSweeny, 2017). Por supuesto, abordar estas cuestiones requerirá un entendimiento al menos básico de la naturaleza de los algoritmos de aprendizaje automático, una de las importantes motivaciones de este artículo.

Otro aspecto importante desde el punto de vista regulatorio es la transparencia. Las empresas están utilizando cada vez más el aprendizaje automático para automatizar de forma importante decisiones que afectan a los consumidores, pero en algunos casos esto puede incrementar la opacidad en comparación con una toma de decisiones humana. Un ejemplo es la decisión de conceder crédito: las entidades financieras aplican algoritmos de aprendizaje automático para decidir qué tipos de préstamos otorgar a qué tipos de clientes, si bien los algoritmos no entienden necesariamente las características clave para predecir el riesgo de reembolso. Los reguladores en ésta y otras situaciones similares tienen un papel que jugar a la hora de asegurar la transparencia y la equidad.

Finalmente, muchos de los datos digitales valiosos para las

aplicaciones de aprendizaje automático están en manos de empresas del sector privado, cuyo principal interés en su explotación es comercial. En la medida en que tales datos también revistan un valor público para el análisis y la formulación de políticas, los reguladores tendrán que habilitar mecanismos para que los datos se transmitan desde las empresas que los recogen a un abanico más amplio de partes interesadas.

NOTAS

(1) Los lectores interesados en una discusión académica más técnica pueden consultar varias excelentes revisiones de la literatura económica (por ejemplo, EINAV y LEVIN, 2014; VARIAN, 2014; MULLAINATHAN y SPIESS, 2017).

(2) Nuestro agradecimiento a BRYAN PARDO de la Northwestern University, el primero en señalar estos puntos al autor.

(3) Algunos de los términos no son palabras inglesas debido a que antes de la estimación los datos se han sometido a *stemming*, o reducción de una palabra a su raíz léxica.

(4) En la siguiente discusión, por «tratamiento» se entiende una variable sobre la que el investigador o responsable de política interviene con el fin de generar un cambio, y un «objeto» significa cualquier variable objetivo en la que se pretende influir.

BIBLIOGRAFÍA

ATHEY, S. (2017), «Beyond prediction: Using big data for policy problems», *Science* 355: 483-485.

ATHEY, S.; BLEI, D.; DONNELLY, R.; RUIZ, F., y T. SCHMIDT (2018), «Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data», *American Economic Review Papers and Proceedings*, próxima publicación.

BAKER, S. R.; BLOOM, N., y S. J. DAVIS (2016), «Measuring Economic Policy Uncertainty», *The Quarterly Journal of Economics*, 131:1593-1636.

BANDIERA, O.; HANSEN, S.; PRAT, A., y R. SADUN (2017), CEO Behavior and Firm Performance, *NBER Working Paper* No. 23248.

BELLONI, A.; CHEN, D.; CHERNOZHUKOV, V., y C. HANSEN (2012), «Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain», *Econometrica* 80: 2369-2429.

BELLONI, A.; CHERNOZHUKOV, V., y C. HANSEN (2014), «High-Dimensional Methods and Inference on Structural and Treatment Effects», *Journal of Economic Perspective*, 28: 29-50.

BERNANKE, B. S.; BOIVIN, J., y P. ELIASZ (2005), «Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach», *Quarterly Journal of Economics*, 120: 387-422.

BLEI, D.; NG, A., y M. JORDAN (2003), «Latent Dirichlet Allocation», *Journal of Machine Learning Research*, 3: 993-1022.

BREIMAN, L. (2001), «Statistical Modeling: The Two Cultures», *Statistical Science*, 16(3): 199-231.

BÜHLMANN, P., y S. VAN DE GEER (2011), *Statistics for High-Dimensional Data: Methods, Theory, and Applications*, Springer Series in Statistics, Springer.

CAVALLO, A., y R. RIGOBON (2016), «The Billion Prices Project: Using Online Prices for Measurement and Research», *Journal of Economic Perspectives*, 30: 151-178.

CROFT, J. (2017, May 4), «Artificial intelligence closes in on the work of junior lawyers», *Financial Times*, retrieved from.

DAVIG, T., y A. SMALTER HALL (2017), «Recession Forecasting Using Bayesian Classification». The Federal Reserve Bank of Kansas City; *Research Working Paper,* 16-06.

EINAV, L., y J. LEVIN (2014), «Economics in the age of big data», *Science*, 346(6210).

EROSHEVA, E. A.; FIENBERG, S. E., y C. JOUTARD (2007), «Describing Disability through Individual-Level Mixture Models for Multivariate Binary Data», *The Annals of Applied Statistics*, 1: 502-537.

GLAESER, E. L.; KIM, H., y M. LUCA (2017), «Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity», *Harvard Business School, Working Paper* 18-022.

GROSS, J. H., y D. MANRIQUE-VALLIER (2014), «A Mixed Membership Approach to the Assessment of Political Ideology from Survey Responses», en Airoldi, E. M., D. BLEI, E. A. EROSHEVA, y S. E. FIENBERG (eds.): *Handbook of Mixed Membership Models and Its Applications*, CRC Press.

HANSEN, S.; MCMAHON, M., y A. PRAT (2018), «Transparency and Deliberation on the FOMC: A Computational Linguistics Approach», *Quarterly Journal of Economics*, próxima publicación.

HARTFORD, J.; LEWIS, G.; LEYTON-BROWN, K., y M. Taddy (2017), «Deep IV: A Flexible Approach for Counterfactual Prediction», *Proceedings of the 34th International Conference on Machine Learning*.

HASTIE, T.; TIBSHIRANI, R., y M. WAINWRIGHT (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations,* Number 143 in Monographs on Statistics and Applied Probability, CRC Press.

HOBERG, G., y G. PHILLIPS (2010), «Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis», *Review of Financial Studies,* 23: 3773-3811.

— (2016), «Text-Based Network Industries and Endogenous Product Differentiation», *Journal of Political Economy,* 124: 1423-1465.

JEAN, N.; BURKE, M.; XIE, M.; DAVIS, W. M.; LOBELL, D. B., y S. ERMON (2016), «Combining satellite imagery and machine learning to predict poverty», *Science,* 353: 790-794.

LEE, D. D., y H. S. SEUNG (1999), «Learning the parts of objects by non-negative matrix factorization», *Nature*, 401: 788-791.

MARR, B. (2015), «Big Data: 20 Mind-Boggling Facts Everyone Must Read», *Forbes,* septiembre https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5741b5117b1e.

MCSWEENY (2017), *Algorithms and Coordinated Effects,* Remarks of Commissioner Terrell McSweeny, University of Oxford Center for Competition Law and Policy, 22 de mayo.

MUELLER, H., y C. RAUH (2017), «Reading Between the Lines: Prediction of Political Violence Using Newspaper Text», *American Political Science Review*, próxima publicación.

MULLAINATHAN, S., y J. SPIESS (2017), «Machine Learning: An Applied Econometric Approach», *Journal of Economic Perspectives*, 31: 87-106.

NG, A. Y., y M. I. JORDAN (2002), «On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes», *Neural Information Processing Systems*.

NIMCZIK, J. S. (2017), *Job Mobility Networks and Endogenous Labor Markets,* no publicado, Humboldt University Berlin.

OCDE (2017), *Algorithms and Collusion: Competition Policy in the Digital Age*, HYPERLINK http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm»www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm

SALCEDO, B. (2015), Pricing Algorithms and Tacit Collusion, no publicado, Pennsylvania State University.

STOCK, J. H., y M. W. WATSON (1999), «Forecasting inflation», *Journal of Monetary Economics*, 44: 293-335.

THORSRUD, L. A. (2016), Words are the new numbers: A newsy coincident index of business cycles, *Working Papers* 4/2016, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

TIBSHIRANI, R. (1996), «Regression Shrinkage and Selection via the LASSO, *Journal of the Royal Statistical Society Series B*, 58: 267-88.

VARIAN, H. R. (2014), «Big Data: New Tricks for Econometrics», *Journal of Economic Perspectives*, 28(2): 3-28.

WIGGLESWORTH, R. (2018), «Can big data revolutionise policymaking by governments?», *Financial Times*, 31 de enero.