XXXII Congreso de la Asociación Latinoamericana de Sociología. Asociación Latinoamericana de Sociología, Lima, 2019.

Big Data, Machine Learning y el poder de la Ciencia de Datos: De su inter- acción con las desigualdades sociales.

Hugo Víctor Claros Haro.

Big Data, Machine Learning y el poder de la Ciencia de Datos: De su interacción con las desigualdades sociales.

Hugo Víctor Claros Haro

Resumen

Los recientes desarrollos de las Ciencias de la Computación y la Estadística, y la expansión de las posibilidades tecnológicas, han convergido en la emergencia de medios de producción, consolidación y análisis de datos sin precedentes en su escala y dinamismo. Estas tecnologías hacen posible el análisis de los perfiles de vida, en sus diversos aspectos, de millones de personas de manera dinámica y con un esfuerzo relativamente trivial para los recursos disponibles para grandes actores.

Sin embargo, la capacidad para producir y usar este tipo de tecnologías no está igualmente distribuido en la sociedad, y por ende los fines que guían tales desarrollos no son neutrales.

La ponencia presenta un balance de las algunas de las interacciones principales de estas tecnologías con las estructuras sociales desiguales de los contextos latinoamericanos.

Palabras clave

Big Data; Machine Learning; Desigualdad.

Introducción

El mundo contemporáneo está cada vez más marcado por el peso que adquiere la generación y manejo de información y conocimiento para el desarrollo de las diversas actividades tanto individuales como colectivas. En ese contexto, la constante evolución de los medios tecnológicos que permiten aquella generación deriva en una constante expansión de la frontera de lo posible y en la emergencia de nuevos y cada vez más complejos retos inmediatos y a largo plazo que interactúan con las condiciones sociales previamente existentes.

La emergencia de tecnologías como la Big data, la Ciencia de datos y el Machine Learning profundiza la velocidad y escala con la que la disponibilidad de tales tecnologías puede influir en la sociedad. Al estar condicionada dicha disponibilidad por los recursos y voluntades de los actores individuales y colectivos preexistentes, los nuevos rumbos

que va adoptando a cada momento este devenir distan de ser neutrales, sino que están marcados por intereses y apuestas, por posiciones y tomas de posición.

Por ello, es importante presentar un balance de algunas de las interacciones principales que asume la disponibilidad de estas tecnologías con las características específicas históricas y sociales del entorno latinoamericano como escenario concreto de desigualdad, de problemas y de posibilidades.

Fundamentación del problema

Por “Big Data” entendemos, como mínimo, conjuntos de datos de gran volumen, velocidad y variedad (Laney, 2001). A estas tres características base, diversos autores e instituciones añaden otras tantas de acuerdo con los factores que les parecen destacables, como la variabilidad, la veracidad, etc. Sin embargo, la discusión detallada de cuán adecuado es o no incluir tales características supera el presente esfuerzo y basta, para dar cuenta de manera introductoria, concentrarse en las tres características universalmente reconocidas como constituyentes de la Big data.

El referirnos al gran volumen de los datos como una esas características centrales es, evidentemente, algo sujeto a cambio y actualización: aquello que ayer se consideraba un volumen grande de datos, puede mañana ser algo considerado manejable, con lo cual la definición se actualizaría. Así, si bien normalmente se asocia el umbral de grandes volúmenes de datos con aquello que está fuera de la capacidad de un procesador normal y de su memoria asociada, dependiendo de quién juzgue las características del conjunto de datos, algunos asociarán el umbral de “gran” conjunto de datos a magnitudes cercanas a los petabytes (Doctorow, 2008).

Por otro lado, entendemos como “Machine Learning” a la situación vinculada a una definición clásica del asunto: consideramos el aprendizaje de un programa de computadora en tanto éste ha aprendido de la experiencia E con respecto a alguna clase de tareas T y una medida de desempeño P, si su desempeño en las tareas en T, medidas a través de P, mejora con la experiencia E. (Mitchell, 1997, p. 2)

De igual forma, entendemos como ciencia de datos a la convergencia entre los desarrollos del campo de la Estadística (Cleveland, 2014) y del área de Ciencias de la Computación. Éste aún es un tópico en constante debate, siendo que cada uno de estos dos campos de origen reclama para sí una mayor cuota de influencia que el otro en esta etiqueta percibida como mixta y compartida. Al igual que las sutilezas de aquello que se

considera Big data, la discusión al respecto supera las posibilidades del presente documento, por lo cual usaremos de manera amplia el término, reconociendo que en su interior operan posibilidades guiadas tanto por el uso de modelos, cuanto por el uso de algoritmos no limitados por la necesidad de demostración matemática.

Por último, la desigualdad en América Latina es un tema de discusión tradicional en la Sociología de la región. Siendo Latinoamérica uno de los territorios más desiguales del planeta, es necesario pensar la emergencia de nuevas tecnologías a la luz de tal desigualdad, más allá de que ésta evolucione e incluso pueda disminuir ligeramente (Galván, Mancero, & Amarante, 2016).

Siendo que las tres tecnologías mencionadas requieren de una dotación de recursos y la elección de unos fines determinados, el presente documento trata de leer sus posibilidades, limitaciones y retos a la luz de aquello que le da contexto a los recursos y elecciones de fines, aquello que hace particular e históricamente específico el rumbo de la adopción de zonas tecnologías en la región.

Metodología

La metodología usada fue la revisión documental de literatura especializada en los campos correspondientes: Estadística, Ciencia de Datos, Machine Learning y estudios sobre desigualdad, así como en otros campos complementarios.

Se intentó considerar la literatura inicial en la que los términos fueron definidos por primera vez, y también literatura que presenta algunos desafíos y condiciones actuales.

A partir de dicha revisión se identificaron algunas de las principales interacciones entre la Big Data, la Ciencia de Datos, el Machine Learning y las estructuras sociales desiguales de la región.

Resultados y discusión

A continuación, se presentan algunas de las principales interacciones identificadas entre las condiciones sociales e históricas preexistentes y la adopción de estas nuevas tecnologías.

Primera interacción principal: las asimetrías de disponibilidad de información se profundizan, exacerbándose las de por sí asimétricas relaciones que establecen los actores en el mercado y la sociedad en general.

En primer lugar, la disponibilidad de información que tienen unos actores sobre otros se hace crecientemente dispar, derivando ello en una desigual posibilidad de anticipar las acciones del resto de actores y de poder situarse frente no sólo a lo que ocurre en el presente, sino de poder leer ese presente como una trayectoria y anticipar escenarios posibles. Esto es importante particularmente en aquello vinculado a las operaciones mediadas por el mercado.

Segunda interacción principal: la producción y propiedad de la información tiende a favorecer a los actores más grandes

Un aspecto fundamental vinculado al anterior es que las posibilidades de producir información no sólo están desigualmente distribuidas en la sociedad, sino que la posibilidad de decidir sobre el uso de la información producida es también desigual. Si bien las diversas leyes que van poco a poco consolidándose en la región establecen, por ejemplo, mecanismos de consentimiento informado a través de los cuales se busca que los individuos tengan que ser previamente puestos al tanto de aquello que están autorizando que se haga con sus datos, en la práctica muchos de los mecanismos a través de los cuales se extrae valor de esos datos y la información generada son opacos al público y pueden funcionar incluso sin el consentimiento explícito de los individuos involucrados (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013; Lipworth, Mason, Kerridge, & Ioannidis, 2017). Un caso típico de esto es el uso de mecanismos como la desanonimización de data a través de matches probabilísticos.

Tercera interacción principal: el valor generado a partir de la información es mayor para los actores con mayores recursos, pues les permite anticipar y tratar de influir

En tercer lugar, el valor generado a partir de la información es mayor para los actores con mayores recursos también debido a que les permite hacer un uso no sólo descriptivo sino predictivo y hasta prescriptivo en función de los recursos que manejen. Por ejemplo, la información producida por actores poseedores de infraestructura de salud, educación, finanzas y otros, les sirve como elementos de influencia corporativa frente al individuo, pero también como bases de construcción argumentativa dirigida a los gobiernos de los estados nacionales.

En este sentido, incluso si se dotara de exactamente la misma información a otro tipo de actores, aquello que les permitiría hacer esta información a estos otros actores con menores recursos de otro tipo sería sin duda algo mucho más modesto que aquello que

les permite hacer actores corporativos consolidados y de múltiples rubros de inversión, en tanto sustenta no sólo sus propias tomas de decisión, sino que les permite generar narrativa al respecto e influir a través de ello. Por ejemplo, es posible presentar el recojo masivo de información como alineado al interés público sin necesariamente compartir con el público las grandes ventajas competitivas que disponer de tal recurso de información brindaría a los actores privados que lo consolidaran (Jurkiewicz, 2018).

Cuarta interacción principal: los estados nacionales tienen también la posibilidad de generar información, pero están limitados por imperativos distintos de los marcados por la lógica que pueden tener otros actores.

Si bien es posible para los estados nacionales implementar ellos también desde sus estructuras de gobierno plataformas orientadas a la generación y uso de información a esta escala (Shahin & Zheng, 2018) y con una potencia comparable de análisis, existe una gran diferencia respecto de los usos privados y las motivaciones privadas que guían tales usos: los gobiernos de los estados nacionales están limitados, al menos teóricamente, por el imperativo de no discriminación y posicionan como anhelo central al bienestar común, a diferencia de las elecciones que puede tomar un actor privado.

Quinta interacción principal: la tensión entre libertad y seguridad se estructura desde desiguales conocimientos sobre qué implica autorizar la producción, almacenamiento y uso de información.

Por último, a la tradicional tensión entre libertad y seguridad, por medio de la cual los estados nacionales y diversos actores privados prometen que los recortes en las libertades individuales a diferentes niveles se harán en aras del incremento de la seguridad anhelada por los individuos, al menos temporalmente, se añade algo importante cuando se habla de estas tecnologías: dado que el solo comprender qué implica autorizar y promover el uso estas tecnologías es un conocimiento desigualmente distribuido en la sociedad (Lipworth et al., 2017), la elección sobre sacrificar libertad en aras de seguridad está también condicionada, ella misma, por la desigualdad preexistente.

Conclusiones

Las salidas a los problemas generados por la adopción de estas nuevas tecnologías en un contexto de tanta desigualdad no necesariamente implican tratar de equiparar las posibilidades de los pequeños actores individuales a aquello que estará a la mano para los grandes actores corporativos, sino que existe también potencial a ser realizado en

la búsqueda de otras formas de generar capacidad de decisión. Por ejemplo, mejorar en la población la capacidad de utilizar data “pequeña” (Onsrud & Campbell, 2007), dado que incluso la disponibilidad de Big Data no necesariamente ha redundado en hallar respuestas creativas y necesarias a los grandes problemas de la sociedad, sino que ha sido posible concentrarla en acometer problemas “pequeños” suficientemente estructurados y sin alto costo por error (Knüsel et al., 2019, p. 196):

Si bien la Big Data amplía los recursos de los que se dispone para responder a la problemática existente, la potencial disponibilidad de estos conjuntos de información no reemplaza ni necesariamente supera las posibilidades brindadas por conjuntos menos complejos de información. Asimismo, muchos de los problemas y deficiencias asociados al actual manejo de tales conjuntos menos complejos de información resultan amplificados al momento de lidiar con Big data y sus tecnologías asociadas.

En ese contexto, es importante que los estados nacionales continúen con sus esfuerzos por generar competencias adecuadas en sus ciudadanos, y es crucial lograr que éstos puedan interpretar mejor los riesgos asociados a las numerosas promesas emergidas de la disponibilidad de estas nuevas tecnologías, especialmente cuando tales promesas estarán condicionadas por los intereses y apuestas de quienes detentan su propiedad y poseen el know how necesario como para utilizar estas nuevas posibilidades como elementos complementarios a sus ejercicios de poder.

Es importante el fortalecimiento masivo de tales competencias en la medida en que la información por sí misma puede ser presentada de múltiples maneras (Matejka & Fitzmaurice, 2017) y detectar qué presentación es más adecuada que otra requiere un nivel mayor de conocimiento, e incluso puede generarse con cada vez mayor facilidad información no auténtica pero suficientemente verosímil (Jia et al., 2018) a partir de la cual los actores pueden tratar de influir en la realidad (por ejemplo, a través de las llamadas “fake news”).

Bibliografía

Cleveland, W. S. (2014). Data science: An action plan for expanding the technical areas of the field of statistics: Technical Areas of the Field of Statistics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *7*(6), 414–417. https://doi.org/10.1002/sam.11239

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, 1376.

Doctorow, C. (2008). Big data: Welcome to the petacentre. *Nature*, *455*(7209), 16–21. https://doi.org/10.1038/455016a

Galván, M., Mancero, X., & Amarante, V. (2016). Desigualdad en América Latina: Una medición global. *Revista de la CEPAL*, *2016*(118), 27–47. https://doi.org/10.18356/ee343975-es

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., … Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 4480–4490). Recuperado de http://papers.nips.cc/paper/7700-transfer-learning-from-speaker-verification-to-multispeaker-text-to-speech-synthesis.pdf

Jurkiewicz, C. L. (2018). Big Data, Big Concerns: Ethics in the Digital Age. *Public Integrity*, *20*(sup1), S46–S59. https://doi.org/10.1080/10999922.2018.1448218

Knüsel, B., Zumwald, M., Baumberger, C., Hirsch Hadorn, G., Fischer, E. M., Bresch, D. N., & Knutti, R. (2019). Applying big data beyond small problems in climate research. *Nature Climate Change*, *9*(3), 196–202. https://doi.org/10.1038/s41558-019-0404-1

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety* (p. 3). Recuperado de https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lipworth, W., Mason, P. H., Kerridge, I., & Ioannidis, J. P. A. (2017). Ethics and Epistemology in Big Data Research. *Journal of Bioethical Inquiry*, *14*(4), 489–500. https://doi.org/10.1007/s11673-017-9771-3

Matejka, J., & Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–1294. https://doi.org/10.1145/3025453.3025912

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Onsrud, H., & Campbell, J. (2007). Big Opportunities in Access to "Small Science" Data. *Data Science Journal*, *6*, OD58–OD66. https://doi.org/10.2481/dsj.6.OD58

Shahin, S., & Zheng, P. (2018). Big Data and the Illusion of Choice: Comparing the Evolution of India's Aadhaar and China's Social Credit System as Technosocial Discourses. *Social Science Computer Review*, 089443931878934. https://bit.ly/2W7Y4cm