Organización Internacional del Trabajo

▶ Nota de la OIT

▶ Nota informativa

Marzo de 2021

Big data y ciencia de datos: casos y experiencias en el ámbito laboral[1]

Walter Sosa Escudero[2]

Principales puntos

- Esta Nota complementa una primera entrega que abordó conceptos, oportunidades y desafíos de *big data* y la ciencia de datos (Sosa Escudero, 2021).
- En esta oportunidad, se revisan algunas implementaciones de *big data* y *machine learning* a cuestiones relacionadas con el mundo del trabajo.
- En particular, en lo que se refiere a estadísticas del mercado laboral, mecanismos de la dinámica del empleo e iniciativas en los sectores público y privado.
- La Nota también se refiere a algunas cuestiones éticas y de transparencia y a la posibilidad de que las estadísticas oficiales interactúen con datos que surgen de *big data* y algoritmos en materia laboral.

El fenómeno de *big data*, entendido como una proliferación de datos masivos producto de interactuar con dispositivos interconectados, afecta a todas las disciplinas y áreas del conocimiento, y, claramente, el mundo del trabajo no es una excepción. Por el contrario, las "huellas digitales" que dejan estas interacciones ofrecen una oportunidad inigualable de aprender acerca de la forma en la que las instituciones y las personas operan, sobre todo en relación a la forma en la que se vinculan y relacionan en el ámbito laboral.

En una nota anterior (Sosa Escudero, 2021) se discuten aspectos generales del nuevo paradigma de datos masivos. En esta nota técnica se discuten algunas implementaciones del paradigma de *big data* y *machine learning* en lo que se refiere a varios aspectos de la dinámica del mercado laboral, incluyendo la interacción entre trabajadores y empleadores, varias iniciativas de los sectores público y privado, la relación entre los nuevos datos y las estadísticas oficiales, el rol de las cuestiones

[1] Nota técnica elaborada para la Organización Internacional del Trabajo. Diciembre de 2020. Se agradecen los valiosos comentarios de Marcela Cabezas, Roxana Mauricio y Fabio Bertranou, y varias interacciones con Laura Ripani, Tomas Castagnino, Carlos Ospino, Leonardo Lucchetti, Wendy Brau y Maria Victoria Anauati.

[2] Profesor plenario de la Universidad de San Andrés. Investigador principal, CONICET (Argentina).

éticas y de transparencia y algunos avances recientes en investigación de frontera.

Antes de comenzar esta revisión, es importante aclarar con más precisión el rol de los algoritmos y *machine learning* en esta visión.

Aspectos preliminares: big data y machine learning

La naturaleza marcadamente interdisciplinar de lo que denominamos como *big data* produjo una rápida proliferación de términos, que posiblemente sea la responsable de las dificultades en dar con una demarcación precisa de su ámbito, a la vez de constituir una auténtica barrera a la entrada a la temática. Como es de esperar, distintos términos muchas veces comparten significados, de modo que resulta complejo diferenciarlos. El problema es exponencialmente mayor cuando se trata de traducirlos al idioma español, toda vez que la literatura asociada se inició en el idioma inglés.

La idea de aprendizaje automático (*machine learning*), numérico o estadístico hace referencia a una variedad de algoritmos computacionales, matemáticos, lógicos y numéricos que permiten aprender a partir de datos. A diferencia de la estadística clásica, en donde se presupone la existencia de una suerte de modelo teórico cuyos parámetros o aspectos desconocidos son estimados a partir de datos, en la lógica de *machine learning* son los propios datos los que guían la construcción de un modelo. Es en este sentido que el método permite "aprender" el modelo, iterativamente, en pos de un objetivo concreto, que en general consiste en predecir, clasificar o tomar decisiones.

A modo de ejemplo, un modelo que clasifica quienes reciben un crédito o no, parte de datos pre-existentes de créditos pasados (otorgados o no) y en base a las características de los receptores (solvencia financiera, cuestiones demográficas, etc.) construye un modelo que predice la probabilidad de repago de un crédito. Este modelo prototípico es evaluado a la luz de nuevos datos y reelaborado para aumentar su precisión predictiva. A este tipo de razonamiento se lo conoce como aprendizaje supervisado: la existencia de datos previos sobre pagos de créditos "guía" la construcción de un modelo para la predicción. En el caso de aprendizaje no supervisado, los datos previos no asisten la clasificación o la predicción. Un ejemplo típico es el uso de una base de datos de características sociodemográficas a fines de clasificar a la población en pobres y no pobres (ver Caruso, Sosa, Escudero y Svarc (2015) para una aplicación de este enfoque). Sin observar el estatus de pobreza de nadie, el objetivo es construir dos grupos, uno con características inferiores que se diferencie del resto, y conjeturar que el primer grupo contiene a los pobres. James, Witten, Hastie y Tibshirani (2103) es una muy buena introducción a los métodos de aprendizaje estadístico. Sosa Escudero (2019) es una introducción informal a estas cuestiones.

Experiencias en el mercado laboral

Pronósticos y nowcasting

Varias fuentes citan al crucial trabajo de "Google Flu Trends" (Ginsberg et al., 2009) como el que da comienzo a una nueva era de la explotación del uso de datos masivos para fines predictivos. Recordemos que dicho trabajo explota las búsquedas de las personas en Google a fines de monitorear y predecir la evolución de la epidemia de Gripe A en 2009. Lazer et al. (2014) discuten varias de las limitaciones de este enfoque. En su momento fue considerado un auténtico logro de *big data*: una conjunción de datos de disponibilidad pública e inmediata y de algoritmos de *machine learning* permite adelantarse casi dos semanas a los guarismos producidos por las oficinas estatales, que dependen de costosísimas encuestas y de un enorme aparato administrativo.

En esa línea, el trabajo de Askitas y Zimmerman (2009) puede considerarse pionero en el uso de big data para medir y predecir la desocupación. Dichos autores siguen una ruta prácticamente idéntica a la del estudio pionero de Google Flu Trends: usar las búsquedas en Google relacionadas con el mercado

laboral a fines de "adelantarse" (*nowcasting*) a las tendencias del mercado de trabajo, otrora captadas a través de encuestas de hogares, muchas veces de cobertura limitada y frecuencia muy baja. La disponibilidad de datos masivos, como las búsquedas en Google, parecen poder resolver ambos problemas: el del rezago informativo y el de la falta de granularidad geográfica. Choi y Varian (2012) y Varian (2009) son buenos surveys de esta literatura iniciática.

Naturalmente, los procesos, fuentes y métodos han avanzado a pasos agigantados en materias de pronósticos, y ya son varios los trabajos que explotan esta innovadora fuente de datos que provee *big data* y las "huellas digitales". A la fecha, posiblemente el más innovador y abarcativo de los trabajos de monitoreo y *nowcasting* sea el de Chetty et al. (2020), que devino en una suerte de "mega proyecto" llamado Economic Tracker. El mismo consiste en un considerable esfuerzo de compatibilización de fuentes de información instantánea, mayormente del sector privado, como datos de usos de tarjetas de crédito, de sitios de ofertas de puestos de trabajo, de movilidad geográfica, entre varios otros. Esta conjunción de datos permite obtener una visión granular (geográficamente, casi a nivel de unas pocas cuadras de diámetro, en varios casos) y virtualmente en tiempo real de la evolución de la economía norteamericana. El trabajo se torna crucial ante la pandemia de la COVID-19, cuyos efectos significativos demandan estrategias rápidas, para las cuales los rezagos del sistema tradicional de monitoreo son inaceptables.

Otro ejemplo reciente de trabajos en esta línea es el que usa información obtenida en tiempo real por la empresa *Burning Glass* y también de las oficinas de seguro de desempleo para estudiar y predecir el efecto de la COVID19 en la demanda laboral. Un punto importante, que reiteraremos en varias ocasiones, es que la irrupción de *big data* llama a una alta interacción entre los sectores público y privado (como Burning Glass o Linkedin), una práctica ya habitual en el sector farmacéutico o agrícola, en donde las empresas del rubro cumplen un rol importante, tanto por su prevalencia en términos de datos como por sus propios intereses en el sector.

> Un caso reciente relacionado con las estadísticas laborales y con la finalidad de revisar el impacto de la COVID-19, lo constituye el modelo de *nowcasting* de la OIT, con el que se ha estado monitoreando la evolución de la pandemia en términos de pérdida de horas de trabajo, para todas las regiones del mundo.

Un caso también reciente, relacionado con las estadísticas laborales y con la finalidad de revisar el impacto de la COVID-19, lo constituye el modelo de *nowcasting* de la OIT, con el que se ha estado monitoreando la evolución de la pandemia en términos de pérdida de horas de trabajo, para todas las regiones del mundo. En esta experiencia se mezclan datos provenientes de encuestas oficiales (encuestas de fuerza de trabajo) y de registros administrativos (empleo registrado) con información proveniente de dispositivos móviles procedentes de los Informes de la Google Community Mobility, así como la data más reciente de Google Trends y los valores más recientes del COVID-19 Government Response Stringency Index (Oxford Stringency Index). Una vez obtenidas la variación de horas de trabajo a partir del modelo, se calculan los equivalentes de puestos de trabajo perdidos a tiempo completo, para lo que se utilizan como referencia las horas semanales trabajadas antes de la crisis de la COVID-19.

Un último caso relevante es el del trabajo de Bailliu et al. (2019), que utiliza datos de diarios en China para monitorear y predecir las condiciones laborales en dicho país. Más concretamente, el trabajo usa sofisticados métodos computacionales para "leer" diarios y noticias, y en base a la intensidad de uso de ciertas palabras, predecir cómo evoluciona el mercado laboral chino. Este tipo de estudio es interesante ya que sugiere que parte de la revolución

de *big data* consiste en alterar lo que se entiende por dato. En el enfoque tradicional, la idea de "dato" era fundamentalmente numérica, sea continua o categórica: datos de desocupación, salarios, experiencia en años, etc. La revolución de *machine learning* hace que dato sea cualquier objeto capaz de ser estudiado sistemáticamente por algoritmos. Desde este punto de vista, un texto, una noticia, una foto, el recorrido de un auto o una imagen satelital son tan "dato" como la serie de desocupación de un país o las respuestas a una encuesta de hogares, en el sentido en que pueden ser estudiadas en forma rigurosa y algorítmica.

La relevancia del sector privado

En la sección anterior adelantamos la idea del rol crucial que cumple el sector privado, tanto como fuente de información relevante como de "motor" en la dinámica del mercado laboral. *Burning Glass* Technologies es una empresa proveedora de servicios de *analytics* para distintos actores del mercado laboral, desde las empresas hasta el sector educativo. Obviamente, Linkedin es un jugador clave, en su doble rol de red social profesional y empresa proveedora de servicios de facilitación en el mercado laboral.

> El programa *Reach for the STAR's, de Accenture*, es un ejemplo de política moderna de avance en el mercado laboral para jóvenes con formación intermedia, que los asiste en sus diseños de sendero laboral a través de un uso inteligente de datos y algoritmos, que detectan "nichos" de productividad que permiten aprovechar sus habilidades latentes.

Las propias empresas del mercado juegan un papel importante, por su impronta y su tamaño. A modo de ejemplo, el programa *Reach for the STAR's*, de *Accenture*, es un ejemplo de política moderna de avance en el mercado laboral para jóvenes con formación intermedia (solo secundario completo), que los asiste en sus diseños de sendero laboral a través de un uso inteligente de datos y algoritmos, que detectan "nichos" de productividad que permiten aprovechar sus habilidades latentes.

En todo caso, estas experiencias sugieren que, en forma acorde con lo que sucede en medicina y agronomía, en el futuro cercano el sector privado cumplirá un rol importante en la investigación aplicada, tanto por su acceso a información importante como por su papel en la dinámica del mercado.

Iniciativas en el sector público

Naturalmente, el sector público, nacional e internacional, cumple un papel fundamental, tanto por su tamaño como por su centralidad en varios ámbitos, fundamentalmente en países en desarrollo. Por "sector público" entendemos tanto a la estructura estatal como a los organismos internacionales.

ESSnet big data es tal vez el proyecto más ambicioso de compatibilización de fuentes modernas y tradicionales, impulsado por el European Statistical System. Se trata de un proyecto a gran escala, que intenta compatibilizar, integrar, explorar y regular la interacción entre los datos masivos de *big data* con el sistema tradicional de estadísticas oficiales. Todavía en una etapa experimental, el programa incluye un módulo relacionado con el mercado de trabajo.

Hay varios proyectos menores, más específicos en relación al uso de *big data*, *machine learning* e inteligencia artificial en la cuestión laboral. Sin exagerar, es posible afirmar que todos los sectores y países (a casi todo nivel de gobierno, desde nacional a municipal) han hecho algún tipo de esfuerzo en adoptar este enfoque, aunque no necesariamente todavía como un proyecto sistemático.

Brookings maneja un proyecto llamado *Visualizing Vulnerable Jobs Across America*, que permite monitorear disponibilidad de trabajos, vulnerabilidad y salarios en regiones muy concretas de los EEUU, casi a nivel de código postal. El BID apoya varias

iniciativas, como el proyecto Bola de Cristal, en Costa Rica o Destino Empleo en Chile, "un servicio digital que busca orientar a las personas que buscan trabajo para la toma de decisiones sobre su futuro laboral", como dice la página web del proyecto, que también involucra al gobierno de Chile y la empresa Movistar. A nivel de gobierno local, Buenos Aires Data es un proyecto iniciado por la Ciudad Autónoma de Buenos Aires, que ofrece a los usuarios acceso a varias fuentes primarias de datos. El Área Metropolitana de Barcelona también ofrece un sistema similar, tal vez con mucha mayor integración con distintas áreas y ejes.

Otras plataformas persiguen un fin tal vez menos operativo, quizás más relacionado con estrategias de gobierno abierto, a fines de transparentar políticas y regulaciones. Ejemplos de este tipo de iniciativas son las implementadas por SERVIR, en Perú, o el de la Dirección General de Servicio Civil de Costa Rica, que presentan detallada información online acerca de las carreras profesionales en los sectores públicos de sus respectivos países.

Investigación y tópicos de frontera

El mercado laboral ocupa un lugar central en la agenda académica de investigación teórica y aplicada en economía y varias ciencias sociales. La limitación para acceder a datos específicos (de índole microeconómica) fue una eterna limitante de estas líneas de investigación, que históricamente dependían de datos de encuestas sistemáticas o de series temporales de frecuencia baja y extensión limitada.

> Big data abre un auténtico "portal" al comportamiento de los agentes económicos en lo referente a sus decisiones laborales.

Big data abre un auténtico "portal" al comportamiento de los agentes económicos en lo referente a sus decisiones laborales.

El estudio de Frey y Osborne (2017) es un ejemplo de este tipo de aproximación moderna. En base a información de *big data* y un meticuloso algoritmo, estiman las chances de que ciertos trabajos sean automatizables. El estudio generó un enorme revuelo en la profesión.

El reciente artículo de Cook et al (2020) explota la copiosa información disponible en la empresa Uber a fines de estudiar las brechas salariales de género. El estudio es no sólo importante por el tema en sí mismo, sino también por el tipo y cantidad de información utilizada: solo en base a los choferes de Uber es posible estudiar el problema de brecha de género con más de un millón de datos distribuidos en el tiempo y geográficamente, tarea impensable hasta hace poco si se utilizara la base de encuestas tradicionales. Este ejemplo muestra el enorme potencial de *big data* en facilitar la ejecución de investigaciones precisas en temas relevantes.

Las contribuciones vienen no solo de la mano de explotar fuentes de datos alternativas sino también de implementar ideas y técnicas del ámbito de *machine learning*. El trabajo de Davis y Heller (2020) usa los más recientes avances en modelos causales de *machine learning* para estudiar el impacto de la implementación de programas de entrenamiento laboral en los jóvenes. Hasta no hace poco, el uso de *machine learning* era fundamentalmente exploratorio, clasificatorio o descriptivo, en claro contraste con los estudios causales que prevalecieron en las dos últimas décadas, basados en experimentos naturales o artificiales. El estudio de Davis y Heller es un ejemplo de que la brecha entre los típicos estudios de *big data-machine learning* (mayormente inductivos y exploratorios) y los de la econometría tradicional (causales) comienza a cerrarse.

Comentarios finales

Si bien todavía incipientes, los estudios basados en *big data* y algoritmos parecen ofrecer una vía promisoria para el ámbito laboral, en varias dimensiones. En primer lugar, como una fuente

alternativa de información, sustitutiva y complementaria, de otras fuentes tradicionales, lo cual plantea una gran oportunidad para la construcción de estadísticas oficiales. En segundo lugar, la interacción de datos masivos y algoritmos abre una alternativa relevante para la propia dinámica del mercado laboral, en particular para el monitoreo y el diseño de políticas para el sector. También se resalta la relevancia de que los sectores público y privado interactúen en estas cuestiones, como es habitual en disciplinas tradicionales como la medicina o la agronomía. Finalmente, a la irrupción inicial de estudios descriptivos y predictivos se suma una saludable tendencia a usar datos masivos y algoritmos para desentrañar fenómenos causales en el ámbito laboral.

Referencias

- Askitas, Nikos & Zimmermann, Klaus F., 2009. "Google Econometrics and Unemployment Forecasting," IZA Discussion Papers 4201, Institute of Labor Economics (IZA).
- Bailliu, Jeannine & Han, Xinfen & Kruger, Mark & Liu, Yu-Hsien & Thanabalasingam, Sri. 2019. Can media and text analytics provide insights into labour market conditions in China?. International Journal of Forecasting. 35. 10
- Caruso, German, Marcela Svarc & Walter Sosa Escudero, Deprivation and the Dimensionality of Welfare: A Variable-Selection Cluster-Analysis Approach, 2015, Review of Income and Wealth, 61, 4, pp. 702-722
- Chetty, Raj, John Friedman, Nathaniel Hendren & Michael Stepner, 2020 The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data, NBER Working Paper No. 26463.
- Choi, H. & Hal Varian, 2012, Predicting the present with Google Trends, Economic Record, 88, 2-9.
- Diamond, Rebecca, Cody Cook, Jonathan Hall, John List & Paul Oyer, 2021, The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers, Review of Economic Studies, en prensa.
- Davis, Jonathan M.V. & Sara B. Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." American Economic Review, 107 (5): 546-50.
- Frey, Carl & Michael A. Osborne, 2017, The future of employment: How susceptible are jobs to computerisation?, Technological Forecasting and Social Change, 114, 254-280.
- Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, 2013, An Introduction to Statistical Learning: with Applications in R. New York, Springer.
- Ginsberg, J., Mohebbi, M., Patel, R. et al., 2009, Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.
- Lazer, D., Kennedy, R., King, G. y Vespignani, A., 2014, The Parable of Google Flu: Traps in Big Data Analysis, Science, 343(6176), pp.1203–1205
- Sosa Escudero, Walter, 2019, Big data, Siglo XXI Editores, Buenos Aires.
- Sosa Escudero, Walter, 2021, Big data y ciencia de datos: conceptos, oportunidades y desafíos, Nota informativa, Organización Internacional del Trabajo. Disponible online en: https://www.ilo.org/wcmsp5/groups/public/---americas/---ro-lima/---sro-santiago/documents/publication/wcms_769307.pdf
- Varian, Hal, 2014, Big Data: New Tricks for Econometrics, Journal of Economic Perspectives, American Economic Association, vol. 28(2), pages 3-28, Spring.

Contacto

Organización Internacional del Trabajo
Oficina de la OIT para el Cono Sur de América Latina.
Santiago de Chile

T: (56-2) 2580-5500
E: santiago@ilo.org
W: ilo.org/santiago