Banco Interamericano de Desarrollo

El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe

Patricio Rodríguez
Norma Palomino
Javier Mondaca

Sector de Conocimiento y Aprendizaje (KNL)
Biblioteca Felipe Herrera (FHL)

RESUMEN DE POLÍTICAS Nº
IDB-PB-266

Julio, 2017

El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe

Patricio Rodríguez
Norma Palomino
Javier Mondaca

Julio, 2017

BID
Banco Interamericano de Desarrollo

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
Rodríguez, Patricio.
El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe / Patricio Rodríguez, Norma Palomino, Javier Mondaca.
p. cm. — (Resumen de políticas del BID ; 266)
Incluye referencias bibliográficas.
1. Big data-Latin America. 2. Big data-Caribbean Area. 3. Political planning-Latin America-Data processing. 4. Political planning-Caribbean Area-Data processing. 5. Public administration-Latin America-Decision making. 6. Public administration-Caribbean Area-Decision making. I. Palomino, Norma. II. Mondaca, Javier. III. Banco Interamericano de Desarrollo. Biblioteca Felipe Herrera. IV. Título. V. Serie.
IDB-PB-266

Códigos JEL: C55
Palabras clave: datos masivos, toma de decisiones, inteligencia de valor público, Big Data, Data Science

http://www.iadb.org

Norma Palomino (npalomino@iadb.org), Banco Interamericano de Desarrollo, Departamento de Conocimiento y Aprendizaje. Patricio Rodriguez y Javier Mondaca, Centro de Investigación Avanzada en Educación, Universidad de Chile.

Resumen

Con el fin de colaborar con la utilización de *Big Data* y *Data Science* en el diseño e implementación de políticas públicas en Latinoamérica y el Caribe, se presentan tres estudios exploratorios realizados por equipos sectoriales del BID en las áreas de productividad a nivel de firma, movilidad urbana sostenible y ciudades inteligentes. Además, se analizan aspectos sensibles del uso del *Big Data* en el marco de políticas públicas, tales como seguridad, propiedad de datos, privacidad y marco ético de uso. Finalmente, se ofrecen recomendaciones para la toma de decisiones y el diseño, implementación y evaluación de políticas públicas por parte de las agencias de gobierno.

Palabras clave: políticas públicas, toma de decisiones, inteligencia de valor público, *Big Data, Data Science*

Índice

1. Introducción

Actualmente, en la actividad económica moderna, los datos se constituyen como un factor esencial para la producción, tal como los activos fijos y el capital humano [1]. Con el advenimiento de las tecnologías de información, los datos han pasado de ser escasos a ser superabundantes [1]–[3]. Esto ha permitido el desarrollo una serie de tecnologías y conocimientos que dependen de estas capacidades [4]. Los posibles impactos positivos abarcan diversos sectores, desde el *retail* y la manufacturación de productos hasta la salud o la administración pública [1].

En este contexto, el objetivo de este documento es revisitar la definición de *'Big Data'* y sus técnicas analíticas en el contexto de la formulación de políticas públicas, con enfoque en América Latina y el Caribe. Lo anterior, a través de la presentación de tres estudios exploratorios realizados por equipos sectoriales del BID en las áreas de movilidad urbana sostenible, ciudades inteligentes y productividad a nivel de firma. Se analizan, además, aspectos sensibles del uso del *Big Data* en el marco de políticas públicas, tales como seguridad, propiedad de datos, privacidad y marco ético de uso. A partir de lo anterior, se introducen recomendaciones, cuya adopción por parte de las agencias de gobierno facilitaría la toma de decisiones fundamentada en datos y evidencia.

El documento se estructura de la siguiente manera; en la sección dos se revisará brevemente el concepto de *'Big Data'*, dando cuenta de las metodologías que utiliza para gestionar los datos, analizarlos, y de las tecnologías empleadas para eso. Luego, en la sección tres, se discutirá, a través de casos exploratorios, cómo aplicar concretamente la analítica sobre datos masivos para generar evidencia que sirva a la toma de decisiones de política pública en América Latina y el Caribe. Finalmente, en la sección cuatro, se entregan conclusiones y recomendaciones.

2. Marco conceptual

El término *'Big Data'* (en adelante, "datos masivos") se ha transformado en una palabra con múltiples acepciones, y en el que se mezclan conceptos relativos tanto a los datos propiamente tales, las tecnologías para su manipulación, las técnicas y tecnologías para su análisis y los profesionales y sus capacidades necesarias para realizar esta tarea. En esta sección, pretendemos arrojar brevemente luces sobre estos aspectos.

2.1. ¿Qué se entiende hasta ahora por datos masivos?

El término datos masivos proviene originalmente del ámbito de las ciencias de la computación, y se refiere a un conjunto de datos cuyo tamaño excede al que puede manejar el software y hardware estándares disponible para capturar, almacenar y analizarlos [1], [4]–[9]. Así, las primeras características de los datos masivos fueron [4], [10]–[14]:

(1) **Volumen:** refiere a la gran cantidad de datos existentes. Asociado a esta característica también están los recursos de almacenamiento y la capacidad de cómputo requeridos para administrar dichos datos.

(2) **Velocidad:** los datos son producidos y analizados a una gran velocidad, en otras palabras, se crean, procesan, analizan y almacenan aceleradamente [8], [13].

(3) **Variedad:** refiere a las fuentes y tipos de datos —texto, audio, video, redes sociales— de datos. Los datos pueden clasificarse en estructurados, semiestructurados e inestructurados [4], [10].

Como esta definición se basa en propiedades técnicas, debe ser revisitada continuamente [15]; razón por la cual se incorporan dimensiones más cualitativas en su caracterización [10], [13], [16], tales como:

(4) **Variabilidad:** cuando el volumen de datos es reducido aparecen observaciones que presentan anomalías respecto de patrones prominentes [17] (usualmente llamados

outliers).En los datos masivos, la cantidad de anomalías puede ser tan abundante que pierden dicha condición, volviéndose parte integrante del fenómeno a analizar (e.g., fenómenos virales en internet [18]).

(5) **Complejidad**: se explica por la múltiple y variada cantidad de fuentes de datos existentes, causada por la proliferación de diferentes dispositivos conectados en línea (e.g., GPS y los sensores del internet de las cosas), datos que pueden ser tanto inter-sujetos e intra-sujeto [19]. La primera se relaciona con la capacidad de recabar datos de muchos sujetos en un instante, mientras que el segundo tipo refiere la capacidad de recabar continuamente datos de un mismo sujeto (e.g., datos biométricos de un sensor de ejercicio).

(6) **Veracidad:** entendida como la calidad, confiabilidad y la certeza asociada a los datos, especialmente en relación a su origen y construcción.

(7) **Representatividad**: cuestiona si los datos masivos representan adecuadamente las poblaciones analizadas, por la naturaleza propia de los datos o los medios establecidos para obtenerlos.

2.2. ¿Cómo se procesa y analizan los datos masivos?

Los datos masivos sin procesar tienen poco valor por sí mismos. Este sólo se obtiene luego de su procesamiento [10] y se relaciona tanto con el retorno a la inversión, como con la posibilidad de construir conocimiento valioso, mejorar procesos, y contribuir a la toma de decisiones disminuyendo la incerteza, entre otros.

Para el procesamiento y análisis de los datos masivos ha surgido una disciplina denominada **Ciencia de Datos** [9], [13]. Esta combina un conjunto amplio de técnicas provenientes de disciplinas tales como Ciencias de la Computación, Matemáticas, Estadística, Econometría e Investigación Operativa [20], [21]. El ciclo de vida del análisis de datos que contempla la Ciencia de Datos se esquematiza en la Figura 1.

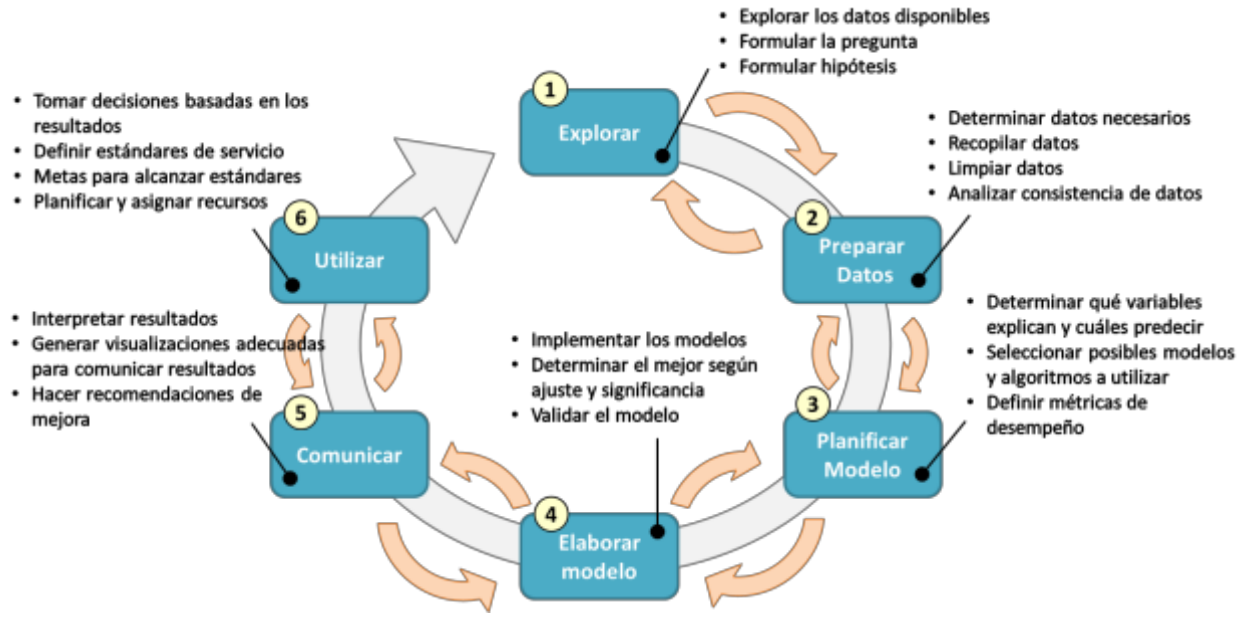

Figura 1: Ciclo de vida del análisis de datos [22].

Como muestra la Figura 1, el ciclo de vida del análisis de datos es un proceso iterativo. Se puede retroceder a etapas previas si se requiere reformular las preguntas en función de la disponibilidad de datos, o reinterpretar los resultados a la luz de nueva evidencia. El procesamiento de los datos masivos se puede sintetizar a partir de dos etapas principales: la gestión de los datos y la analítica de datos [10], [23].

La **gestión de los datos** se compone de tres aspectos: (1) adquisición y almacenamiento de los datos, (2) limpieza y depuración de los datos y, (3) la preparación para su análisis. La **analítica de datos** se refiere a la respuesta de preguntas y/o hipótesis formuladas a partir de técnicas de modelamiento y análisis.

Este ciclo de análisis requiere de profesionales especialistas con una formación sólida en alguna de las Ciencias de la Computación, uso y desarrollo de aplicaciones, modelamiento, estadística, analítica y matemáticas. A estos profesionales se les denomina **Científicos de Datos**, quienes exploran, generan preguntas, realizan análisis de escenarios y cuestionan los supuestos y procesos existentes utilizando múltiples fuentes de datos de diferentes orígenes [13], [24].

3. Uso de analítica avanzada para la toma de decisiones, el diseño, implementación y evaluación de políticas públicas

En el caso de la toma de decisiones, el diseño, implementación y evaluación de políticas públicas el objetivo del uso de la Ciencia de Datos es producir evidencia que sea pertinente, de calidad y oportuna, para así fundamentar y orientar decisiones. Esto significa diagnosticar problemas que pasan inadvertidos o desapercibidos y, por lo tanto, son imposibles de accionar [25]. Este proceso se denomina "toma de decisiones guiadas por datos" (del inglés *'Data-driven decision making'*) [9]. Así el análisis de datos masivos puede ser utilizado para la mejora de la administración pública, a través de la generación de más y mejores soluciones que satisfagan necesidades de salud, educación, transporte, vivienda, atención e inclusión de grupos desaventajados, entre otras, a partir de contextos sociales, demográficos y territoriales particulares.

Por ejemplo, en el primer *'Big Data Innovation Challenge'* del Banco Mundial [26], se premió una serie de iniciativas al servicio de la generación de valor público utilizando datos masivos en distintos ámbitos; entre los que se encuentran:

- **Pobreza**: en India se utilizaron imágenes satelitales nocturnas para analizar la cobertura eléctrica en las más de 600.000 aldeas del país. Estas imágenes se utilizaron en Sri Lanka y Pakistán para generar modelos de predicción de pobreza. Por su parte, en Nigeria, se evaluó la relación entre pobreza y baja eficacia de mercado, combinando los precios mensuales de cientos de *'commodities'* con imágenes satelitales nocturnas (que dan cuenta de sectores con electricidad).

- **Crimen y seguridad**: En Colombia, se usó la información de las rutas del sistema de buses (*'Bus rapid transit', BRT)*, en conjunto con el modelamiento de terreno riesgoso *('risk terra in modeling'* en inglés) para buscar asociación entre crimen e infraestructura urbana. De esta manera, se asociaron ciertas zonas y horarios a mayor ocurrencia de delitos.

- **Transporte**: en Filipinas se desarrollaron las aplicaciones *OpenRoad* y *Open Traffic*. La primera es un portal interactivo que permite a los usuarios hacer seguimiento a proyectos viales con financiamiento público, y entregar *feedback*. La segunda, permite visualizar y analizar información de la velocidad del tráfico, utilizando información entregada por taxistas (GPS e información de teléfono celular). En Bielorrusia se desarrolló otra aplicación llamada *RoadLab*, que permite evaluar la calidad de la superficie de las calles y caminos utilizando el acelerómetro de los teléfonos celulares y entregando su posición a través del GPS.

- **Salud**: en Sudáfrica se usaron técnicas de datos masivos para identificar a nivel nacional, provincial, distrital o de dependencia de salud, los lugares en que se atienden las mayores proporciones de pacientes con SIDA [27].

A continuación, con el fin de ilustrar los principales desafíos, oportunidades y aprendizajes de la analítica basada en datos, se revisarán tres casos de implementación en América Latina y el Caribe.

3.1. Movilidad Urbana Sostenible, Datos Masivos y Políticas Públicas: estudio de la movilidad de los ciclistas en la Ciudad de Rosario, Argentina

Descripción del caso, necesidades y/o problemáticas detectadas

En esta ciudad se estudió la movilidad de los ciclistas (utilizando dispositivos de georreferenciación) para entender: 1) los patrones de movilidad de los ciclistas en relación con la infraestructura vial existente (sean ciclovías o no), 2) la relación de estos patrones y los accidentes de tránsito, y 3) las posibles mejoras de la infraestructura vial [28].

El estudio tuvo dos fases. En la primera, se instalaron GPS en 150 bicicletas del sistema público "Mi bici, tu bici", con un 85% de cobertura. En la segunda, se instalaron dispositivos de localización GPS en sus bicicletas a 40 usuarios voluntarios privados, para complementar la información de las bicicletas públicas ya que su patrón de movilidad no está geográficamente limitado por la ubicación de los terminales de bicicletas públicas; permitiendo evaluar otro tipo de recorridos.

Se recabaron datos respecto a la velocidad y el recorrido de los viajes públicos y privados. Para el análisis de los accidentes, se recopilaron datos de distintas agencias públicas (Observatorio Vial de la Agencia Provincial de Seguridad Vial de Santa Fe; y el Ente de la Movilidad de Rosario). Finalmente, se realizaron entrevistas focales a ciclistas privados para validar la información cuantitativa obtenida.

Metodologías para analizar los datos

Los datos obtenidos desde fuentes privadas y públicas fueron recolectados en un periodo de dos semanas y seis semanas respectivamente. En ambos casos, sólo se consideraron datos de días hábiles. Posteriormente, se sistematizó la cantidad de viajes, los tiempos invertidos en los mismos, las distancias o la velocidad promedio, y los ejes viales más utilizados. Esta información se resumió a nivel de mes, día, horario o individuo y diferenciada según rango etario y género.

Para procesar los datos se exploraron visualmente creando mapas que mostraban los principales corredores utilizados por los ciclistas, la velocidad del tránsito de bicicletas, los focos de accidentes y la gravedad de dichos siniestros. Así, se identificaron vías, cruces o zonas específicas que, por la frecuencia y gravedad de los accidentes, requerían mayor atención. Para complementar estos resultados, se hicieron entrevistas a ciclistas para profundizar en estos focos y/o en problemas generales.

El estudio detectó que los ciclistas prefieren las calles que cuentan con ciclovías, por ser más rápidas y seguras. El análisis de datos identificó focos de accidentes en las calles sin ciclovías (pero con altos volúmenes de ciclistas), y ejes viales sin ciclovías que son altamente utilizados por los ciclistas en sus desplazamientos. Dichos resultados informan la toma de decisiones para mejorar la infraestructura pública. En este caso específico, el estudio da cuenta de la necesidad de mejoras de la calle Bulevar Oroño, que cumple una función de eje articulador de la ciudad y es ampliamente utilizada por ciclistas.

3.2. Computando una nueva trayectoria para la gobernanza: innovaciones en datos masivos en América Latina y el Caribe

Descripción del caso, necesidades y/o problemáticas detectadas

Se analizaron cuatro casos de ciudades inteligentes (*smart cities* en inglés) en Latinoamérica. Con ciudades inteligentes, se consideran tres aspectos relativos al gobierno de una ciudad: transparencia, eficiencia e innovación continua [29].

Bahía Blanca (Argentina): la iniciativa '¿Qué pasa Bahía Blanca?' (QPBB) surge de la tensión entre activistas medioambientales y la industria petroquímica de la zona. El gobierno instaló sensores que miden continuamente efluentes líquidos y la contaminación acústica de origen

industrial en las vecindades de las plantas petroquímicas, y en algunas de ellas también se dispuso de cámaras que las monitorean en vivo. La información así producida se publicó en una plataforma y aplicación móvil, permitiendo seguir en tiempo real la contaminación del aire y acústica generada y que además quedó disponible como datos abiertos.

Córdoba (Argentina): para incentivar el uso del transporte público y la movilidad, se desarrolló un sistema de seguimiento de la flota del transporte público. Gracias a la instalación de dispositivos GPS en todos los buses del sistema público[1], se pudo recopilar información respecto a tiempos de viaje, recorridos e ingresos. Además, se recolectó información del pago a través del terminal utilizado por 'tarjeta única' (sistema de pago unificado para el sistema de transporte público de la ciudad) tal como datos sobre la hora, tarifa y localización de cada transacción realizada a través de la tarjeta.

São Bernardo do Campo (Brasil): como respuesta a los problemas del crecimiento de la ciudad –tales como congestión, acceso desigual a los servicios y seguridad ciudadana–, el gobierno federal realizó un esfuerzo progresivo por mejorar la infraestructura y la logística detrás de diversos servicios públicos. Para esto creó *Você SBC*, una aplicación móvil que permite enviar quejas de los ciudadanos y sugerencias relacionadas con una amplia gama de servicios no-urgentes. Esta aplicación permite conocer y hacer seguimiento de las necesidades de la ciudad y sus habitantes.

Fortaleza (Brasil): el proyecto 'Fortaleza Inteligente' consta de tres proyectos pilotos: el primero utiliza información del sistema GPS en los buses del transporte público, para evitar retrasos y sobrepaso del límite de su capacidad. El segundo utiliza datos del sistema de bicicletas públicas de la ciudad para analizar su uso y generar evidencia para su expansión. El tercero es un *dashboard* que unifica indicadores del sistema de transporte en su conjunto, entregando visualizaciones en la web.

Metodologías usadas para analizar los datos

Para el análisis de las experiencias señaladas se utilizó el 'modelo de maduración de datos masivos urbanos' propuesto por el BID [29]; éste consiste en una rúbrica que considera cinco dimensiones: (1) datos abiertos, (2) cultivar ecosistemas de datos, (3) analítica, (4) toma de decisiones basadas en datos, y (5) participación y servicios públicos. Cada una de estas dimensiones tiene una rúbrica de adopción que va desde resolver problemas específicos del momento, hasta un nivel en que la apropiación está completa y la mejora continua se releva como importante. Esto permite estimar en qué nivel de desarrollo (o madurez) se encuentra la iniciativa de datos masivos urbanos evaluada.

3.3. Utilizando datos a nivel de empresa para estudiar el crecimiento y dispersión en el factor de productividad total

Descripción del caso, necesidades y/o problemáticas detectadas

En este estudio se tomaron datos a nivel de empresa y se estimó el crecimiento y dispersión en la productividad total de los factores (*total factor productivity* en inglés, o TFP). La TFP es la proporción de la producción que no está explicada por las cantidades de insumos que necesita para ser producida; y su nivel estará determinado por cuán eficiente e intensivamente se usan los insumos en la producción [30].

El estudio analizado calculó la TFP para cerca de 20 millones de empresas, durante ocho años en alrededor de 30 países, entregando un panorama más general respecto al estado y evolución de dicho indicador. Para dar cuenta del amplio escenario económico, se requirió recolectar una cantidad importante de datos de empresas. Estos datos fueron obtenidos desde la base de datos Orbis [31] que dispone de toda la información necesaria para computar las TFP, incluyendo balances financieros y estimaciones de las medidas de productividad.

[1] Incluyendo los que ya contaban con otro GPS instalado por el operador respectivo.

Metodologías para analizar los datos

En términos metodológicos, el proyecto contó con dos grandes fases: la preparación de los datos y el cálculo de funciones de productividad. El gran desafío fue la limpieza y preparación de los datos, la que requirió de una cantidad considerable de tiempo, además del capital humano adecuado para realizarla.

Posteriormente, se calcularon las funciones de producción, a partir modelos de regresión que determinaron cuánto se produjo a partir de los insumos disponibles. Para esto se emplearon cuatro metodologías distintas, todas basadas en mínimos cuadrados ordinarios, para salvaguardar la robustez de los resultados [32]. En términos computacionales, esta fase fue la parte más intensa del proceso, tanto por la cantidad de datos, como por la variedad de metodologías utilizadas en los cálculos. Luego se calcularon las elasticidades de los factores de producción, que son el peso de cada factor de producción en su industria, país respectivo, e incluso a nivel de firma.

En términos de resultados, se observa que la dispersión identificada no tiene una relación clara con la combinación entre TFP promedio (considerando los distintos rubros) y países. Respecto al crecimiento, al controlar por los niveles de línea base de los TFP, se observa que casi en todas las medidas existe una relación negativa con el crecimiento futuro y las TFP.

4. Discusión

Es importante recordar que los datos masivos y la Ciencia de Datos, como toda nueva herramienta, tiene ciertas limitaciones metodológicas (como por ejemplo su representatividad) y cuestionamientos relativos a la privacidad, usos éticos, legales, intelectuales y de seguridad, que es necesario tomar en consideración.

4.1. Desafíos y limitaciones

Análisis de datos, metodologías y tecnologías

Con respecto a la metodología, se ha considerado erróneamente la superabundancia de los datos como sinónimo de representatividad. Esto puede observarse en el tercer caso presentado, donde, si bien se dispone de abundantes datos, éstos provienen en su mayoría de firmas europeas. Otros ejemplos pueden ser cuando se recopilan datos por canales digitales, estos sólo son representativos de ciertos usuarios más activos y, en el mejor de los casos, sólo de aquellos que tienen acceso a tecnologías de información y comunicación, que en Latinoamérica y el Caribe aún está lejos del 100% [33]. De esta manera, los fenómenos de baja y sobre-representación, y multiplicidad presentan claros desafíos a la capacidad de realizar inferencias generalizables, ya que cuestionan si los datos masivos representa la diversidad de la población bajo estudio [34]. Por consiguiente, tanto los aspectos metodológicos como la confiabilidad de las fuentes son especialmente relevantes [4], [35].

Por otro lado, si bien gran parte del procesamiento de los datos puede automatizarse gracias a la existencia de diversas tecnologías, esto no significa que los científicos de datos no deban tomar una serie de decisiones que puedan ser discutibles o arbitrarias. Por ejemplo en el ciclo de vida del análisis de datos (ver Figura 1), a la hora de implementar procesos de extracción y limpieza de datos, pueden preferirse ciertos tipos de análisis en desmedro de otros y además cometerse errores al interpretar los resultados [35]. Esto significa que la analítica de datos masivos no es una disciplina enteramente **objetiva**, y tiene un componente importante de **subjetividad** [35].

En lo concerniente al fracaso de algoritmos predictivos, este no se debe necesariamente a los datos mismos, sino a los errores que se comenten en el análisis e interpretación [4]. Por ejemplo, Es altamente posible que el gran volumen de los datos produzca correlaciones espurias entre variables y alta significancia estadística de los resultados [4], [9], [36], siendo un potencial problema el

"sobreajuste"[2] de los modelos, y su potencial generalización. Esto porque los datos y, por lo tanto su significado, siguen siendo dependientes del contexto [35].

En conclusión, no se puede esperar que el uso de analítica avanzada de datos masivos sustituya por sí misma a métodos de investigación y análisis más tradicionales, sino que, por el contrario, debe y puede servir de complemento para otros [37], especialmente de índole cualitativa. En otras palabras, la analítica avanzada sobre datos masivos se constituye como una herramienta y no un fin en sí mismo.

Privacidad, aspectos éticos y legales, seguridad y pertenencia

El hecho que los datos masivos puedan contener una gran cantidad de variables personales hace necesaria la consideración de consideraciones éticas y legales respecto a

(1) La **protección de la información privada de las personas**, es decir, mantener el anonimato de las personas cuyos datos se están analizando.

(2) El **análisis de los datos privados**, es decir, el cuestionamiento acerca de la inequidad o perjuicio que genera la intromisión en la vida privada[3] de las personas [39], [40].

(3) La **propiedad de los datos masivos** y los correspondientes derechos y licencias para su administración, mantenimiento, explotación y uso.

(4) La **responsabilidad legal** identifica quién es responsable cuando las acciones de análisis de datos masivos generan consecuencias negativas: aspectos de pertenencia y protección de datos, privacidad de las personas y protección del consumidor, problemas con la seguridad de los datos entre otros [41].

En este contexto, es importante que, cuando los gobiernos se apoyen en entidades externas (universidades o empresas) para sus análisis, se resguarde adecuadamente la propiedad de los datos, se establezcan mecanismos para su protección, y una prohibición de uso posterior para otros fines. Ese tipo de temas legales necesitan de pronta clarificación [1] y son parte ineludible de una definición de grandes datos para la formulación de políticas públicas.

4.2. Recomendaciones

De esta forma, es posible desarrollar una "**inteligencia de valor público**" (un equivalente social de la "inteligencia de negocios") que tiene la potencialidad de ser un componente estratégico para la toma de decisiones y el diseño, implementación y evaluación de políticas públicas.

Sobre la adopción de la inteligencia de valor público en las agencias de gobierno

A continuación, se proponen una serie de recomendaciones para la adopción de la inteligencia de valor público en las agencias de gobierno. En primer lugar, para implementar este tipo de proyectos se requiere de una serie de **capacidades institucionales** dentro del gobierno. Algunos autores identifican a lo menos tres dimensiones: capital humano, tecnología y desarrollo de estrategias [20], [25], [42].

(1) **Capital humano**: para cubrir tareas como estudiar y pensar respecto de la información disponible; limpiar, preparar, formatear y asegurar la confiabilidad de los datos y realizar capacitación específica en el análisis de datos y soluciones basadas en ellas. Asimismo, para que las iniciativas sean sostenibles en el tiempo, se requiere que el sector público desarrolle su

[2] Del inglés "*overfitting*", corresponde cuando un modelo estadístico se ajusta demasiado a los datos de entrenamiento, reduciendo su validez predictiva fuera de dicho conjunto de datos.

[3] Un ejemplo de esto último es el del supermercado estadounidense Target, que, en función de la compra de ciertos productos, predice qué clientes están embarazadas y les envía promociones acordes. El problema ocurrió cuando una adolescente embarazada, que no lo había hablado con sus padres, recibió una de estas promociones [38].

propio capital humano. Es necesario, además, formar 'consumidores inteligentes' que analicen críticamente la información.

(2) **Tecnología**: recursos tecnológicos para el uso de grandes conjuntos de datos y los servicios de software y almacenamiento asociados a ellos. Igualmente, se apunta a generar interoperabilidad entre sistemas de distintas agencias y/o departamentos, y herramientas para generar acciones a partir de los datos.

(3) **Desarrollo de estrategias**: un plan que determine qué preguntas son urgentes contestar, qué datos recopilar y con qué técnicas analizarlos. Además de alianzas estratégicas con organizaciones cuya misión sea apoyar el uso de datos, recursos calidad y confiabilidad de la información disponible.

En consecuencia, establecer una **institucionalidad** es importante para mantener las iniciativas en el tiempo [29]. Además, esta institucionalidad debe hacerse cargo de lograr una comunicación **transparente y fluida con otras entidades externas**, esto significa compartir los datos entre sus distintas agencias [14] y desarrollar un liderazgo dentro del gobierno para establecer cómo se usarán los datos masivos y para qué. También es relevante que los propios tomadores de decisiones se involucren, permitiendo un acceso oportuno a los datos (sobre todo a aquellos con una utilidad limitada a ciertos periodos de tiempo), y facilitando el cambio cultural hacia procesos de toma de decisiones basadas en estos.

Transparentar la analítica utilizada para generar la evidencia

Otro aspecto necesario a considerar es la **rendición de cuentas ante la ciudadanía**. Se debe **documentar y transparentar los procesos de análisis llevados a cabo** para que sean auditables y respondan a los mecanismos de rendición de cuentas. De esta forma, existe la oportunidad para la mejora continua de los análisis y sus resultados, la diseminación dentro del sector público de las metodologías empleadas y—sobre todo— la posibilidad de corregir errores a tiempo. Especialmente cuando existen filtraciones de la información privada de las personas, o se producen inequidades producto de las recomendaciones erradas de un algoritmo.

Respecto al **capital humano,** son necesarios tanto los profesionales que llevan a cabo el proceso de análisis como los "**consumidores inteligentes**" de la evidencia producida. La labor de estos últimos es la formulación de preguntas y el análisis crítico de la información recibida, cuestionando las fuentes, supuestos y metodología utilizados para producirlos [25].

Respecto a los **científicos de datos**, se requieren profesionales con competencias bastante específicas para llevar a cabo análisis que produzcan información valiosa y que alimente el ciclo de decisiones guiadas por datos. Sin embargo, estos profesionales son escasos. Tampoco las dificultades enfrentadas en torno a la capacidad profesional se acaban solucionada la escasez. La utilización progresiva de datos relativos a la conducta humana se aleja del área de conocimiento dentro de la ingeniería o las llamadas ciencias duras. Esta situación demanda de un análisis trabajo multidisciplinario y multisectorial, en diferentes contextos sociales, demográficos y geográficos. En otras palabras, se necesitan profesionales que puedan desenvolverse competentemente en esa diversidad.

4.3. Oportunidades

Nivel de desarrollo (o madurez) de proyectos de datos masivos y de los "consumidores inteligentes" de evidencia basada en análisis de datos masivos

Una primera oportunidad es la utilización de la rúbrica desarrollada por Townsend & Zambrano-Barragan [29] para evaluar las iniciativas de relacionadas con datos masivos. Este es un instrumento que, con algunas adaptaciones, puede utilizarse para evaluar la madurez general de cualquier proyecto de análisis de datos masivos dentro del sector público, y de este modo, identificar las capacidades, potencialidades y debilidades de los proyectos, en pos de su mejora.

Del mismo modo, se puede tener en consideración la rúbrica sobre “consumidores inteligentes” [43] para asegurar un nivel mínimo de competencias para lograr hacer interpretación y uso de la evidencia producida a partir del análisis de datos masivos.

Compartir y diseminar datos dentro del sistema público

Una segunda oportunidad es que los proyectos de analítica pública generen una serie de datos con un uso potencial más allá de la iniciativa que los produjo. De esta manera, se podrían presentar claras oportunidades de sinergia entre distintas agencias de gobierno con el potencial de desarrollar análisis que den cuenta de la necesidad de toma de decisiones y desarrollo de políticas multisectoriales, como por ejemplo el caso del transporte, la contaminación ambiental y la concentración de centros educativos [44]. Esto requiere avanzar en políticas de creación y curación continua de los datos generados, y el desarrollo de una institucionalidad que: 1) lidere el uso de análisis de datos masivos para generar una cultura de toma de decisiones basadas en evidencia; 2) se haga cargo de la sustentabilidad de mantener y administrar los datos masivos, manteniendo todos los resguardos necesarios, y 3) promueva la comunicación clara y fluida con otras agencias gubernamentales y entidades externas (como universidades y centros de investigación).

Tipos de problemática a abordar

Una tercera familia de oportunidades es un tipo específico de problemas de “predicción pura” [45], para los que no se necesitan establecer una causalidad para apoyar la toma de decisiones. Por ejemplo: ¿es la posibilidad de lluvia lo suficientemente alta para salir con paraguas? En dicho caso, no se necesita saber qué causa la lluvia, solo estimar si se producirá o no. De esta forma, se pueden emplear técnicas de aprendizaje automático supervisado, utilizando datos históricos para entrenar un algoritmo que entregue un pronóstico de mejor calidad y más oportuno que el que podría realizar un experto humano. Así, por ejemplo, se han generado aplicaciones para: estimar automáticamente el nivel socioeconómico de una cierta zona geográfica en base a información satelital para establecer políticas de ayuda social en un territorio [46]; estimar el riesgo de deserción escolar de un alumno [47], [48] y escoger qué intervenciones es más costo-efectiva para retenerlo [49], y para la mejora de las políticas de fiscalización usando inspecciones predictivas basadas en reseñas en línea de clientes [50].

Referencias Bibliográficas

[1] J. Manyika *et al.*, "Big Data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011.

[2] Staff Science, "Challenges and Opportunities", *Science*, vol. 331, nº 6018, pp. 692–693, nov. 2011.

[3] The Economist, "Data, data everywhere", *The Economist*, feb-2010.

[4] S. T. McAbee, R. S. Landis, y M. I. Burke, "Inductive reasoning: The promise of Big Data", *Hum. Resour. Manag. Rev.*, 2016.

[5] J. Chen *et al.*, "Big Data challenge: a data management perspective", *Front. Comput. Sci.*, vol. 7, nº 2, pp. 157–164, abr. 2013.

[6] L. Manovich, "Trending: The promises and the challenges of big social data", *Debates Digit. Humanit.*, vol. 2, pp. 460–475, 2011.

[7] M. R. Parks, "Big Data in communication research: Its contents and discontents", *J. Commun.*, vol. 64, nº 2, pp. 355–360, Abril 2014.

[8] D. J. Power, "Using 'Big Data' for analytics and decision support", *J. Decis. Syst.*, vol. 23, nº 2, pp. 222–228, Abril 2014.

[9] F. Provost y T. Fawcett, "Data science and its relationship to big data and data-driven decision making", *Big Data*, vol. 1, nº 1, pp. 51–59, 2013.

[10] A. Gandomi y M. Haider, "Beyond the hype: Big Data concepts, methods, and analytics", *Int. J. Inf. Manag.*, vol. 35, nº 2, pp. 137–144, Abril 2015.

[11] R. Kitchin, "Big data and human geography: Opportunities, challenges and risks", *Dialogues Hum. Geogr.*, vol. 3, nº 3, pp. 262–267, 2013.

[12] R. Kitchin, "Big Data, new epistemologies and paradigm shifts", *Big Data Soc.*, vol. 1, nº 1, pp. 1–12, 2014.

[13] I.-Y. Song y Y. Zhu, "Big Data and data science: what should we teach?", *Expert Syst.*, vol. 33, nº 4, pp. 364–373, Agosto 2016.

[14] L. Tomar, W. Guicheney, H. Kyarisiima, y T. Zimani, "Big Data in the public sector: Selected applications and lessons learned", Inter-American Development Bank, 2016.

[15] L. Hill, F. Levy, V. Kundra, B. Laki, y J. Smith, *Data-Driven Innovation for Growth and Well-being*. OECD, 2014.

[16] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, y G. Alor-Hernández, "A general perspective of Big Data: applications, tools, challenges and trends", *J. Supercomput.*, vol. 72, pp. 3073–3113, 2015.

[17] B. F. Welles, "On minorities and outliers: The case for making Big Data small", *Big Data Soc.*, vol. 1, nº 1, pp. 1–2, 2014.

[18] K. Nahon y J. Hemsley, *Going viral*, 1ª ed. Polity Press, 2013.

[19] J. D. Morrison y J. D. Abraham, "Reasons for enthusiasm and caution regarding Big Data in applied selection research", *Ind. Psychol.*, vol. 52, nº 3, pp. 134–139, 2015.

[20] J. A. Marsh, J. F. Pane, y L. S. Hamilton, "Making Sense of Data-Driven Decision Making in Education", 2006. [En línea]. Disponible en: http://www.rand.org/pubs/occasional_papers/OP170.html. [Accedido: 28-ene-2017].

[21] UNESCO, "Policy brief - Learning Analytics". UNESCO Institute for Information Technologies in Education, 2012.

[22] B. Schmarzo, *Big Data: Understanding How Data Powers Big Business*. Wiley, 2013.

[23] A. Labrinidis y H. V. Jagadish, "Challenges and Opportunities with Big Data", *Proc. VLDB Endow.*, vol. 5, nº 12, pp. 2032–2033, 2012.

[24] IBM, "What is a Data Scientist? – Bringing big data to the enterprise", 2015. [En línea]. Disponible en: http://www-01.ibm.com/software/data/infosphere/data-scientist/?cm_mc_uid=23497733502814413056500&cm_mc_sid_50200000=1441305650. [Accedido: 07-sep-2015].

[25] M. Bienkowski, M. Feng, y B. Means, "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics". U.S. Department of Education, Office of Educational Technology, 2012.

[26] K. M. Kelm *et al.*, "Big data innovation challenge : pioneering approaches to data-driven development", The World Bank, 107751, ene. 2016.

[27] W. MacLeod, J. Bor, K. Crawford, y S. Carmona, "Analysis of Big Data for better targeting of ART adherence strategies : spatial clustering analysis of viral load suppression by South African province, district, sub-district and facility (April 2014-March 2015)", The World Bank, 2015.

[28] H. Terraza, P. Deregibus, C. Galeota, y M. Ponce de León, "Movilidad urbana sostenible, datos masivos y políticas públicas: estudio de la movilidad de los ciclistas en la ciudad de Rosario (Argentina) a través del uso de dispositivos de geo-referenciación", Banco Interamericano de Desarrollo, 2016.

[29] A. Townsend y P. Zambrano-Barragan, "Computing a new trajectory for urban gobernance: Urban Big Data innovation in Latin America and the Caribbean", Banco Interamericano de Desarrollo, 2016.

[30] D. Comin, "Economic Growth", en *Economic Growth*, S. N. Durlauf y L. E. Blume, Eds. Palgrave Macmillan UK, 2010, pp. 260–263.

[31] Bureau van Dijk, "Orbis | Detailed global private company information", 2017. [En línea]. Disponible en: http://www.bvdinfo.com/en-gb/our-products/company-information/international-products/orbis. [Accedido: 12-ene-2017].

[32] D. Bahar, "Using firm-level data to study growth and dispersion in total factor productivity", The Brookings Institution Harvard Center for International Development, 2016.

[33] CEPAL, "Estado de la banda ancha en América Latina y el Caribe 2016", CEPAL, sep. 2016.

[34] F. Kreuter y R. D. Peng, "Extracting Information from Big Data: Issues of Measurement, Inference and Linkage", en *Privacy, Big Data, and the Public Good*, J. Lane, V. Stodden, S. Bender, y H. Nissenbaum, Eds. Cambridge University Press, 2014, pp. 257–275.

[35] D. Boyd y K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon", *Inf. Commun. Soc.*, vol. 15, nº 5, pp. 662–679, 2012.

[36] K. Crawford, "The hidden biases in Big Data", *Harvard Business Review*, 01-abr-2013.

[37] D. Lazer, R. Kennedy, G. King, y A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, vol. 343, nº 6176, pp. 1203–1205, mar. 2014.

[38] L. Floridi, "Big data and their epistemological challenge", *Philos. Technol.*, pp. 1–3, 2012.

[39] C. Miller, "When Algorithms Discriminate", *The New York Times*, 09-jul-2015.

[40] Nature, "More accountability for big-data algorithms", *Nat. News*, vol. 537, nº 7621, p. 449, sep. 2016.

[41] European Big Data Value Partnership, "European Big Data value strategic research & innovation agenda: Version 0.99", jul. 2014.

[42] B. Means, C. Padilla, A. DeBarger, y M. Bakia, *Implementing Data-Informed Decision Making in Schools: Teacher Access, Supports and Use*. US Department of Education, 2009.

[43] P. Rodríguez, N. Palomino, y J. Mondaca, “El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe”, Inter-American Development Bank, Discussion Papers & Presentations, may 2017.

[44] P. Rodríguez *et al.*, “Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile”, 2016.

[45] J. Kleinberg, J. Ludwig, S. Mullainathan, y Z. Obermeyer, “Prediction policy problems”, *Am. Econ. Rev.*, vol. 105, nº 5, pp. 491–495, 2015.

[46] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, y S. Ermon, “Combining satellite imagery and machine learning to predict poverty”, *Science*, vol. 353, nº 6301, pp. 790–794, ago. 2016.

[47] C. Escobar y F. Lolas, “Desarrollo de un sistema prototipo para la detección temprana de la deserción escolar en escuelas públicas chilenas"”, Memoria de Título, Universidad Adolfo Ibáñez, Santiago de Chile, 2015.

[48] Mineduc, “Informe de Piloto de Modelo Predictivo, Seguimiento de Estrategias de Apoyo (Sistema de Alerta Temprana)”. jul-2015.

[49] Microsoft, “Predicting student dropout risks, increasing graduation rates with cloud analytics”, ago-2016. [En línea]. Disponible en: https://customers.microsoft.com/en-us/story/tacomapublicschoolsstory. [Accedido: 14-dic-2016].

[50] J. S. Kang, P. Kuznetsova, M. Luca, y Y. Choi, “Where not to eat? Improving public policy by predicting hygiene inspections using online reviews.”, en *EMNLP*, 2013, pp. 1443–1448.