COLOMBIA POTENCIA DE LA VIDA

DANE
DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICA
70 AÑOS

Revisión de REFERENTES INTERNACIONALES

DIRECCIÓN DE REGULACIÓN, PLANEACIÓN, ESTANDARIZACIÓN Y NORMALIZACIÓN (DIRPEN)

REVISIÓN DE REFERENTES INTERNACIONALES

(1) Avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC)

(2) Metodología para calcular el indicador ODS 10.7.3: número de personas que murieron o desaparecieron en el proceso de migración hacia un destino internacional

(3) Desafíos en la producción de estadísticas oficiales con nuevos métodos de recolección de datos

(4) Diez propuestas sobre el aprendizaje automático en la estadística oficial

Febrero 2024

CONTENIDO

Lista de tablas

Introducción

Este reporte tiene el propósito de apoyar el conocimiento, la generación de recomendaciones y propiciar acciones acordes a las necesidades de temáticas líderes del Departamento Administrativo Nacional de Estadísticas (DANE) y del Sistema Estadística Nacional (SEN), a partir de una revisión prospectiva que involucra referentes internacionales de diferente naturaleza y el rol en el ecosistema de datos, incluyendo oficinas nacionales de estadística, organizaciones no gubernamentales e institutos de investigación, etc.

Con ello, se busca enriquecer los trabajos que se vienen desarrollando al interior de las áreas técnicas del DANE y las instancias de coordinación del SEN, considerados prioritarios en concordancia con el Plan Estratégico Institucional y las agendas de trabajo e investigación.

Para tal fin, la revisión de referentes constituye una investigación prospectiva de la práctica internacional, en función del tema de análisis, de las organizaciones mencionadas anteriormente. Los temas que se abordan en cada reporte se priorizan al considerar la urgencia de la necesidad, a partir de una lista de temas construida de la consulta directa realizada a los directivos DANE, los directores técnicos y los coordinadores de las mesas estadísticas del SEN. La profundidad y el detalle de las revisiones está asociada a las preguntas clave, las perspectivas, el alcance y la disponibilidad de información, que pretende dar una respuesta adecuada y generar valor.

En esta versión del reporte se abordan cuatro temas: avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC); metodología para calcular el indicador ODS 10.7.3: número de personas que murieron o desaparecieron en el proceso de migración hacia un destino internacional; desafíos en la producción de estadísticas oficiales con nuevos métodos de recolección de datos, y diez propuestas sobre el aprendizaje automático en la estadística oficial.

Revisión de
REFERENTES INTERNACIONALES

1.

Avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC)

1. Avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC)

1.1. Resumen

Los aportes y las contribuciones de los ciudadanos a los datos son esenciales para la superación de brechas y retos que afrontan las estadísticas en la actualidad. Este tipo de datos, se definen según el Departamento de Asuntos Económicos y Sociales de las Naciones Unidas[1] como la participación de los ciudadanos en múltiples procesos de la cadena de valor, los cuales garantizan que los procesos de toma de decisiones sean más inclusivos, equitativos, abiertos y transparentes al contemplar aspectos propios y particulares de las comunidades, de igual forma a través de estos se accede a información de difícil recolección, como aquella vinculada a comunidades rurales o en territorios de laborioso ingreso.

Sin embargo, estos datos se ven enfrentados a grandes desafíos, riesgos y retos como lo manifestó el Foro Económico Mundial para 2024[2] quienes mencionan que las falencias en los datos son un problema global, es por esto que desde la Agenda 2030 se ha motivado a actores estatales, miembros de la sociedad civil, organismos multilaterales y entidades privadas a generar estrategias de colaboración conjunta que permita el intercambio de experiencias, la generación de agendas de trabajo y el desarrollo de recomendaciones, marcos regulatorios, guías y otros documentos, que permitan establecer referentes para la toma de decisiones, es por ello que surge la necesidad de conocer los avances de los países entorno a este aspecto.

Dado el auge de los proyectos de datos generados por los usuarios y con el fin de brindar herramientas a las naciones para trabajar proyectos con este tipo de información, Global Partnership for Sustainable Development Data ha generado una guía para la interacción con datos generados por los usuarios[3], la cual aborda los flujos de trabajo para generar datos, la participación y la idoneidad de los datos para su uso final y ha servido como insumo de muchos países para la generación de sus documentos propios, de igual forma desde Naciones Unidas se desarrolla actualmente el marco de Copenhague sobre datos generados por la ciudadanía, un documento referente que define los posibles tipos de datos generados por los ciudadanos y ofrece un entendimiento común de los conceptos relevantes.

[1] Disponible en: https://unstats.un.org/UNSDWebsite/citizen-data/

[2] Disponible en: https://es.weforum.org/agenda/2024/01/estos-son-los-mayores-riesgos-globales-a-los-que-nos-enfrentamos-en-2024-y-mas-alla/

[3] Disponible en: https://www.data4sdgs.org/resources/choosing-and-engaging-citizen-generated-data-guide

En Colombia, los datos generados por la ciudadanía son de vital importancia, especialmente dada la relevancia de la participación de la sociedad civil en el proceso de la producción estadística más allá de la recolección de datos y la necesidad de integrar la información proveniente de fuentes campesinas e indígenas en las estadísticas oficiales, es por esto que en la actualidad y como parte del cumplimiento de la estrategia 5 , acción 15, meta 35 del Plan Estadístico Nacional (PEN) 2023 – 2027, la cual menciona que se debe "elaborar un documento de recomendaciones, buenas prácticas, directrices, mecanismos y alianzas para la participación ciudadana en la generación, la validación y el uso de datos estadísticos"[4].

1.2. Síntesis de hallazgos

A continuación, en la Tabla 1 se presenta una breve descripción de los principales hallazgos de la revisión de referentes internacionales sobre avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC).

Tabla 1. Principales hallazgos

Referente	Avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC)
Uganda	El gobierno de Uganda, a través de su oficina nacional de estadística, ha desarrollado una guía para la generación de datos por parte de los ciudadanos que funciona como un kit de herramientas que incluye gestión de garantía de calidad de la información, el monitoreo y la evaluación.
Kenia	"Datos generados por ciudadanos en Kenia: una guía práctica para comprender y generar datos de calidad" se establece como un recurso esencial respaldado por la Oficina Nacional de Estadísticas (KNBS) y está fortalecido por la colaboración de las organizaciones de la sociedad civil. Desde las mejores prácticas hasta la privacidad de los datos, la guía establece un marco integral para mejorar la calidad y la utilización de la información generada por ciudadanos.
Ghana	En Ghana ha habido un progreso significativo en la generación y el uso de Datos Generados por la Ciudadanía (DGC). Dichos progresos se encuentran reflejados en documentos, investigaciones y guías entre otros. Asimismo, están regulados por marcos políticos como la Ley de Protección de Datos de Ghana (2012) y la Ley de Libertad de Información de Ghana (2019). A pesar del progreso significativo, aún existen desafíos para la adopción y el uso de DGC en Ghana; estos incluyen resolver

[4] Disponible en: https://www.sen.gov.co/sites/default/files/noticias-files/2023-12/plan_estadistico_naciona_2023-2027_0.pdf

Referente	Avances, investigaciones, guías y marcos regulatorios sobre Datos Generados por la Ciudadanía (DGC)
	la falta de conocimiento social y capacidad de uso de mecanismos y herramientas, mejorar una infraestructura que actualmente es deficiente y eliminar la desconfianza entre ciudadanos y gobierno. Estos desafíos deben abordarse para garantizar que los DGC se utilicen de manera efectiva para el desarrollo del país.
Portal de datos europeo	Portal de datos europeo desarrolló el informe Data.europa.eu and Citizen-generated Data, cuyo principal objetivo es proporcionar una visión de los DGC que pueden formar parte de portales de datos abiertos y cómo incluirlos por parte de las administraciones públicas.
Canadá	Para el caso canadiense no se encontraron documentos con lineamientos sobre el aprovechamiento de datos generados por la ciudadanía, pero se encontraron experiencias donde se evidencia su aplicación. Se identificaron tres modalidades: el crowdsourcing, el uso de datos abiertos y las encuestas en articulación con la ciudadanía.
Perú	El Gobierno de Perú desarrolló el Modelo de Datos Abiertos Gubernamentales. Este modelo considera un enfoque de procesos y cadena de valor de los datos abiertos, con el propósito de tener una visión completa y compartida de todos los procesos de nivel estratégico, operativo y apoyo o soporte, incluyendo los procesos de medición del desempeño e impacto. Tiene como usuarios todas aquellas personas que tienen la capacidad de transformar los datos abiertos gubernamentales en nuevos productos y servicios públicos. Estos datos pueden ser transformados, combinados o relacionados para facilitar la generación de nuevos recursos de información como vistas, tablas dinámicas, gráficos y mapas de fácil interpretación para los ciudadanos.
OCDE	La Organización para la Cooperación y el Desarrollo Económico (OCDE) reconoce la importancia de los DGC para mejorar la toma de decisiones, la transparencia y la participación pública. En este contexto, la OCDE ha desarrollado diversas iniciativas: facilitar la generación y el acceso a los DGC, fomentar el uso de los DGC para la innovación y abordar los desafíos relacionados con los DGC.

Fuente: DANE a partir de las revisiones de referentes.

1.3. Revisión de referentes

En esta sección se presentará de forma sintetizada la revisión de referentes internacionales.

1.3.1. Uganda

En adopción de la resolución de la ONU sobre la necesidad de avanzar en la generación y el uso de datos complementarios, Uganda desarrollo un kit de herramientas de datos generados por el ciudadano[5]. Lo anterior, teniendo en cuenta que las fuentes de datos tradicionales del país indican que Uganda cuenta con gran cantidad de información de este tipo, especialmente en temas relacionados con igualdad de género y empoderamiento de las mujeres. Este documento surge como respuesta a la necesidad de los ciudadanos por datos confiables en los cuales se vean reflejadas las preocupaciones de la comunidad y se fortalezcan los procesos de promoción de las iniciativas internas y la rendición de cuentas, según Chris N. Mukiza (PhD), director ejecutivo de la oficina de estadísticas de Uganda[6].

Esta guía otorga un enfoque estándar para que los productores de datos de fuentes no tradicionales, actuales y potenciales implementen durante la recopilación, el análisis y la difusión de estadísticas. La guía está dirigida a organizaciones de la sociedad civil e instituciones del sector privado, especialmente, aquellas que implementan programas relacionados con igualdad de género y empoderamiento de las mujeres. Los usuarios del kit de herramientas podrán compilar sistemáticamente datos confiables para respaldar la toma de decisiones, el seguimiento y la evaluación basados en evidencia dado que el documento proporciona pasos aceptables y fáciles de seguir para producir datos generados por la ciudadanía y cuya calidad pueda evaluarse con respecto a los estándares de Uganda para este fin, previa a su proclamación como estadísticas oficiales. Esta metodología, contribuye a garantizar que las voces de los ciudadanos, especialmente los grupos y los individuos marginados sean escuchadas y proporciona elementos básicos para coordinar y gobernar el ecosistema de datos con nuevos actores.

El conjunto de herramientas generado por Uganda se encuentra orientado a las particularidades del país sustentando prácticas y ejemplos locales y se constituye como un estándar que todos los productores potenciales de datos generados por la ciudadanía deben adoptar para producir datos confiables, utilizables y accesibles. Sin embargo, la oficina de estadística indica que las organizaciones tienen la libertad de elegir las herramientas que sean factibles dentro de sus posibilidades sin comprometer la objetividad de los datos.

Algunos de los beneficios que menciona la oficina estadística de este kit de herramientas es la mejora de las habilidades y las competencias en la recopilación de los datos generados por la ciudadanía, dado

[5] Disponible en: https://africa.unwomen.org/sites/default/files/Field%20Office%20Africa/Attachments/Publications/2021/12/UgandaCGD2910202102.pdf

[6] Disponible en: https://africa.unwomen.org/sites/default/files/Field%20Office%20Africa/Attachments/Publications/2021/12/UgandaCGD2910202102.pdf

que se llegan a acuerdos para la recopilación, el procesamiento, el análisis, la presentación de informes y la difusión sistemática de la información. De igual forma, esta guía, aumenta la disponibilidad de datos generados por los ciudadanos de calidad y eleva el perfil público de este tipo de información como fuente confiable.

El conjunto de herramientas posee un enfoque diferencial de género, por lo cual, aborda en primer lugar los indicadores de los Objetivos de Desarrollo Sostenible (ODS) relacionados con este aspecto. Posteriormente, describe el marco legal que cobija los datos generados por la ciudadanía en Uganda, vincula la caja de herramientas con otros marcos y políticas estadísticas de la nación, presenta una conceptualización con relación a las fuentes de datos tradicionales y no tradicionales y sitúa un concepto estándar sobre datos generados por los ciudadanos. De igual forma, reconoce algunas ventajas y limitaciones que enfrentan los datos y presenta la cadena de valor de estos, establece el marco de gestión de la calidad de los datos generados por la ciudadanía con relación a la igualdad de género y el empoderamiento de las mujeres y establece criterios para realizar procesos de monitoreo y evaluación.

Finalmente, presenta el marco legal para la transversalización de los datos generados por la ciudadanía y su transformación en estadísticas oficiales, sin dejar de lado el enfoque de género propio de la guía.

1.3.2. Kenia

Datos generados por ciudadanos en Kenia: una guía práctica para comprender y generar datos de calidad[7]. La guía sobre datos generados por ciudadanos en Kenia se crea por la Global Partnership como un recurso, para asistir a los usuarios en la comprensión y la generación de datos de calidad provenientes de ciudadanos. Su finalidad es mejorar los procesos de toma de decisiones y la formulación de políticas. La estructura y el contenido de la guía se presentan a continuación, suministrando una visión integral de su enfoque.

Introducción: la introducción sienta las bases al abordar el panorama de los DGC y presenta estadísticas oficiales en Kenia para establecer un contexto para comprender el papel de los DGC.

Mejores prácticas: este segmento de la guía se dedica a exponer las mejores prácticas para la producción de datos de calidad, extrayendo conocimientos valiosos de las estadísticas oficiales. Sirve como guía para orientar los procesos de generación de datos y asegurar la calidad y la fiabilidad de la información obtenida.

[7] Disponible en: https://www.data4sdgs.org/sites/default/files/file_uploads/Citizen-Generated%20Data%20Improving%20Quality%20and%20Use%20for%20Policy%20and%20Decision-making%20in%20Kenya.pdf

Cadena de valor de datos: la guía desglosa la cadena de valor de los datos, desde la identificación de necesidades hasta la evaluación, ofreciendo dirección en cada etapa del proceso de generación de datos. Este enfoque integral busca maximizar la utilidad y la relevancia de los datos generados por ciudadanos.

Privacidad de datos: en este apartado destaca la privacidad de los datos como una prioridad clave en la producción de datos y resalta la importancia de salvaguardar la información personal.

Recursos adicionales: para enriquecer la comprensión y el conocimiento sobre los datos generados por ciudadanos, la guía incluye al final de cada capítulo material de lectura adicional. Estos recursos adicionales actúan como herramientas complementarias para mejorar la competencia y la práctica en el manejo de datos ciudadanos.

Anexo: la guía contiene un apartado llamado anexo que abarca una lista de verificación, definiciones de términos estadísticos clave y una hoja de ruta para institucionalizar los DGC entre las Organizaciones de la Sociedad Civil (OSC) y las Oficinas Nacionales de Estadística (ONE). Esta sección proporciona un marco estructurado para la implementación y seguimiento de los DGC.

La guía incorpora conocimientos de estándares y principios internacionales y nacionales para estadísticas oficiales y no oficiales, incluidos los Principios Fundamentales de las Estadísticas Oficiales de las Naciones Unidas y el Marco Nacional de Garantía de Calidad de las Naciones Unidas. También refleja el marco legal para las estadísticas en Kenia y estudios de caso del trabajo del DGC implementado a nivel local e internacional.

El desarrollo de la guía implicó talleres de cocreación, encuestas, entrevistas a informantes clave y reuniones con diversas partes interesadas, incluidas organizaciones de la sociedad civil, la Oficina Nacional de Estadísticas de Kenia (KNBS) y organizaciones internacionales. La guía tiene como objetivo proporcionar orientación práctica y herramientas para mejorar la calidad y el uso de los datos generados por los ciudadanos en Kenia, contribuyendo en última instancia a procesos de toma de decisiones más informados y basados en evidencia.

La Oficina Nacional de Estadísticas de Kenia (KNBS) desempeña un papel crucial en respaldar el uso de datos generados por ciudadanos al proporcionar aportes técnicos, incluyendo experiencia en el diseño de instrumentos de estudio para garantizar la confiabilidad de las herramientas de recopilación de datos. La KNBS respalda iniciativas de capacitación para enumeradores y capacitadores, asegurando que los procesos de recopilación cumplan con estándares establecidos. Además, desempeña un papel en la garantía de calidad, asegurando que los datos generados cumplan con estándares documentados, promoviendo la confiabilidad y la comparabilidad de los datos. Colabora activamente con diversas partes interesadas, como organizaciones de la sociedad civil y otros productores de datos, compartiendo lecciones y metodologías para mejorar la eficacia y la eficiencia de las iniciativas de datos generados por ciudadanos.

Las OSC en Kenia desempeñan un papel fundamental en la contribución a la generación de datos de calidad generados por los ciudadanos mediante diversas estrategias. En primer lugar, facilitan la participación ciudadana al interactuar con las comunidades al sensibilizar sobre la importancia de los datos generados por los ciudadanos y empoderando a las personas para compartir sus experiencias y perspectivas. Asimismo, las OSC ofrecen iniciativas de capacitación para dotar a los ciudadanos de habilidades y herramientas necesarias para recopilar, analizar y reportar datos de manera efectiva, garantizando así la calidad y la confiabilidad de la información. Igualmente, organizan iniciativas de presentación de informes ciudadanos, encuestas y debates de grupos focales para recopilar información específica, aprovechando su conocimiento y redes para llegar a una audiencia más amplia. Establecen mecanismos para verificar y validar la exactitud y la confiabilidad de los datos generados por los ciudadanos y mantener altos estándares de calidad. Además, apoyan el reconocimiento y la utilización de estos datos en los procesos de toma de decisiones y políticas, destacando su valor como fuente complementaria. Las OSC colaboran con agencias gubernamentales, instituciones académicas y otras partes interesadas, fortaleciendo así la generación y la utilización de datos generados por ciudadanos en un enfoque multisectorial.

El uso de datos generados por ciudadanos en procesos de toma de decisiones y políticas presenta beneficios fundamentales. Primero, promueve la inclusividad y el empoderamiento al permitir la incorporación de diversas perspectivas ciudadanas en las decisiones al otorgar a los ciudadanos un papel activo en la configuración de políticas que los impactan. Además, ofrece oportunidad y granularidad al proporcionar información detallada y en tiempo real sobre problemas locales, permitiendo una comprensión más actualizada y específica de las necesidades comunitarias. Los datos generados por ciudadanos informan políticas adaptadas al capturar opiniones y experiencias directas al responder de esa manera las necesidades y las prioridades reales de la población.

1.3.3. Ghana

Ghana presenta un progreso significativo en la generación y el uso de DGC. A continuación, se presenta un resumen de algunos de los avances clave en las áreas de documentos, investigaciones, guías, marcos regulatorios e información relacionada:

Documentos

- Política Nacional de Datos Abiertos de Ghana (2020): esta política reconoce la importancia de los DGC y establece un marco para su apertura y uso[8].
- Estrategia Nacional de Datos de Ghana (2021): esta estrategia describe cómo Ghana aprovechará los datos, incluidos los DGC, para promover el desarrollo económico y social[9].

Investigaciones

- Estudio sobre el Potencial de los Datos Generados por los Ciudadanos para el Desarrollo de Ghana (2021): este estudio encontró que los DGC pueden tener un impacto significativo en una variedad de sectores, incluyendo la agricultura, la salud, la educación y la gobernabilidad.

- Investigación sobre las Barreras para la Adopción de Datos Generados por los Ciudadanos en Ghana (2022): esta investigación identificó una serie de barreras para la adopción de DGC, como la falta de conocimiento, la infraestructura y la confianza.

Guías

- Guía para la Protección de Datos Personales en Ghana (2020): esta guía proporciona orientación sobre cómo proteger los datos personales, incluidos los DGC.

- Guía para la Ética del Uso de Datos Generados por los Ciudadanos (2022): esta guía proporciona principios éticos para la recopilación, el uso y la difusión de DGC.

Marcos regulatorios

- Ley de Protección de Datos de Ghana (2012): esta ley establece un marco para la protección de datos personales, incluidos los DGC[10].

[8] Disponible en: https://data.gov.gh/

[9] Disponible en: https://data.gov.gh/about

[10] Disponible en: https://www.dataprotection.org.gh/resources/downloads/data-protection-act

- Ley de Libertad de Información de Ghana (2019): esta ley permite a los ciudadanos acceder a información en poder de entidades públicas, incluidos los DGC[11].

Ejemplos de uso de DGC en Ghana

- Monitoreo de la calidad del agua: los ciudadanos utilizan sensores para recopilar datos sobre la calidad del agua en los ríos y los lagos.

- Seguimiento de la salud: los ciudadanos utilizan aplicaciones móviles para registrar su estado de salud y compartir datos con los profesionales de la salud.

- Rendición de cuentas del gobierno: los ciudadanos utilizan datos para monitorear el desempeño del gobierno y exigir transparencia.

Desafíos: a pesar del progreso significativo aún existen desafíos para la adopción y el uso de DGC en Ghana. Estos incluyen:

- Falta de conocimiento y capacidad: muchas personas no saben qué son los DGC ni cómo usarlos.

- Infraestructura deficiente: la infraestructura de datos en Ghana es deficiente, lo que dificulta la recopilación, el almacenamiento y el análisis de DGC.

- Falta de confianza: existe una falta de confianza entre los ciudadanos y el gobierno, lo que dificulta que los ciudadanos compartan sus datos.

Ghana ha hecho un progreso significativo en la generación y el uso de DGC. Aún hay desafíos que abordar para garantizar que los DGC se utilicen eficazmente para el desarrollo del país.

1.3.4. Portal de datos europeos

Desde el Ministerio de Transformación Digital de España se hace referencia al informe que publica el Portal de Datos Europeo, el cual está enfocado en los DGC.

[11] Disponible en: https://www.dataprotection.org.gh/resources/downloads/conference

En la actualidad existe una escasez de este tipo de datos dentro de los portales de datos abiertos europeos causado por la falta de publicación y gestión de DGC por parte de las administraciones públicas.

En el documento se realiza un análisis de diversos portales de datos abiertos, cuyo principal objetivo es proporcionar una visión de los DGC que pueden formar parte de dichos portales y cómo incluirlos por parte de las administraciones públicas. Cabe destacar que durante el análisis se establece un marco para la descripción, la referencia y la caracterización de los DGC[12].

Según este informe, los ciudadanos residentes o los que visitan Europa generan grandes cantidades de datos al momento de realizar sus actividades diarias. La generación y la recolección de estos datos se puede realizar como resultado de una elección consciente o explícita de los ciudadanos. Como ejemplos de esto se puede mencionar cuando los ciudadanos participan en recursos públicos como Wikidata o OpenStreetMap, cuando envían un reclamo o sugerencia al sitio web de su municipio o cuando contribuyen o seleccionan datos para iniciativas de ciencia ciudadana.

La captura de datos puede ser inconsciente o implícita, es decir, sin necesariamente ser consciente de que se generan y almacenan, a pesar de haber aceptado las condiciones generales de protección de datos o las cookies en los sitios web. Esto ocurre, por ejemplo, cuando se registra la geolocalización de los ciudadanos mientras se desplazan con su teléfono móvil o cuando escanean su tarjeta de transporte al entrar en el transporte público[13].

Gran parte de este DGC es gestionado por empresas privadas. Los ciudadanos dan su consentimiento para que se recopilen dichos datos, ya que normalmente se benefician de los servicios de uso gratuito ofrecidos por estas organizaciones. A su vez, estas organizaciones obtienen el derecho de utilizar dichos datos para mejorar sus servicios o proporcionar otros servicios de valor añadido o productos de datos basados en dichos datos (por ejemplo, calificaciones de empresas y datos de ocupación en Google Places y FourSquare, informes de FixMyStreet sobre grafitis y baches, etc.). El DGC también puede ser gestionado por fundaciones y asociaciones sin ánimo de lucro, con un objetivo de bien público (por ejemplo, Wikipedia, Wikidata, OpenStreetMap). Y, en última instancia, podrán ser gestionados por las administraciones públicas[14].

En este informe se plantea que los DGC no están ampliamente representadas en los portales de datos abiertos principalmente porque los portales de datos abiertos se han limitado a publicar datos directamente generados y gestionados por las administraciones públicas y es en este punto donde el

[12] Disponible en: https://datos.gob.es/es/documentacion/dataeuropaeu-y-los-datos-generados-por-ciudadanos

[13] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

[14] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

portal de datos europeos a través de este informe plantea que es un buen momento para reinvestigar las oportunidades y los retos que puede traer la inclusión de este tipo de DGC en portales de datos abiertos.

Dado que en la actualidad los conjuntos de datos existentes generados por los ciudadanos tienen menos impacto, mientras que los portales y los espacios pierden a un grupo de partes interesadas fundamental al no incluirlos.

El análisis de este informe se centra en DGC que no contiene datos personales o en los que los datos personales se anonimizan o agregan convenientemente.

Con la implementación de los DGC, por parte de las entidades públicas y las organizaciones privadas, se complementan los datos que ya tienen y pueden ofrecer una mejor cobertura y calidad con menores costos, en comparación con las metodologías tradicionales y menos descentralizadas de recopilación de datos.

Fuera del alcance de este informe estará el análisis de los datos que las personas producen cuando utilizan servicios digitales propiedad de empresas (por ejemplo, Foursquare, Waze y Citymapper) que mantienen conjuntos de datos privados de colaboración abierta y los utilizan de acuerdo con el consentimiento de los usuarios[15].

La metodología que se utilizó consta de tres pasos:

1. Selección de fuentes bibliográficas, en el que se seleccionaron artículos relevantes para el dominio DGC.

2. Diseño de un marco de análisis, en este paso se revisaron los artículos en detalle para derivar el marco de análisis.

3. Análisis de campo de portales de datos gubernamentales abiertos.

1.3.5. Canadá

Para el caso canadiense no se encontraron documentos con lineamientos sobre el aprovechamiento de datos generados por la ciudadanía, pero se encontraron experiencias donde se evidencia su aplicación. Se identificaron tres modalidades: el crowdsourcing, el uso de datos abiertos y las encuestas en articulación con la ciudadanía.

En una presentación de Statistics Canada en el Centro Interuniversitario de Estadísticas Sociales de Quebec se definió *crowdsourcing* como un método de recolección no probabilístico donde se invitó a

[15] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

todos los miembros de un segmento de población a participar voluntariamente en ejercicios de recolección de datos en un tema de interés[16]. El *crowdsourcing* implica recolectar información de una gran comunidad de usuarios y se fundamenta en el principio de que los ciudadanos individuales son expertos dentro de sus ambientes locales y sirve para validar otras fuentes de datos complementarios para asegurar la calidad de los resultados[17]. En una presentación con expertos de las Naciones Unidas, se pudo evidenciar que Statistics Canada reconoce el *crowdsourcing* como una metodología de captura de datos generados por la ciudadanía[18].

Algunos proyectos que han recopilado información mediante *crowdsourcing* son Crowdsourcing Cannabis, donde se indagaba por el precio del cannabis antes de su legalización[19], Medición de la Canasta de Mercado (para la medición de la línea de pobreza)[20] y las encuestas de impactos del COVID-19[21].

Por otro lado, el sitio web de Gobierno Abierto de Canadá, define a los datos abiertos como datos comprensibles para las máquinas que pueden ser libremente usados, reutilizados y distribuidos por cualquiera, sujetos solamente, como máximo a citar y distribuir de forma similar[22]. Statistics Canada considera en su sitio web que el uso de datos abiertos invita a la innovación, no solo mediante canales gubernamentales sino también apoyado en organizaciones de base, individuos y negocios[23].

Statistics Canada hace uso de datos abiertos con el proyecto "Open Database of Buildings" (ODB), que centraliza y armoniza un repositorio de información geoestadística relacionada con la huella de ciertas edificaciones edificios, construido a partir de 65 conjuntos de datos abiertos provenientes de diferentes fuentes de datos abiertos de divisiones administrativas del gobierno canadiense[24].

El ODB tiene una particularidad que vale la pena mencionar, y es que se origina de los resultados tomados de un proyecto de estadística experimental de *crowdsourcing* de datos. En 2016 se utilizó la plataforma de datos abiertos OpenStreetMap y se invitó a posibles contribuyentes a georreferenciar de

[16] Disponible en: https://www.ciqss.org/sites/default/files/documents/Crowdsourcing.pdf

[17] Disponible en: https://www.statcan.gc.ca/en/our-data/where/crowdsourcing

[18] Disponible en: https://unstats.un.org/sdgs/files/meetings/harnessing-data-by-citizens-for-public-policy-and-SDG-monitoring/Session2a-2-Canada_citizen_generated_data_C.Williams.pdf

[19] Disponible en: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5263

[20] Disponible en: https://communitydata.ca/sites/default/files/Understanding-MBM-STC-Feb12.pdf

[21] Disponible en: https://www.ciqss.org/sites/default/files/documents/Crowdsourcing.pdf

[22] Disponible en: https://open.canada.ca/en/frequently-asked-questions

[23] Disponible en: https://www.statcan.gc.ca/en/our-data/where/open-data

[24] Disponible en: https://www.statcan.gc.ca/en/lode/databases/odb

manera voluntaria información sobre edificios en la ciudad de Gatineau, Ottawa[25]. Posteriormente se integró esta información con otros conjuntos de datos disponibles en una base de datos que incluye los datos transversales a todas las fuentes de información.

Por último, el desarrollo de encuestas de manera articulada con la ciudadanía se ha implementado con éxito cuando la participación de la comunidad es crucial para el éxito de la operación estadística. Se puede citar como ejemplo el conteo de población sin hogar "Everyone Counts"[26], donde se hizo el recuento de la población sin hogar de 66 comunidades con el apoyo de equipos locales compuestos por organizaciones dedicadas a brindar asistencia a este segmento de población, equipos de trabajo municipales, servicios de emergencia y voluntarios que se tomaron el tiempo de interactuar con las personas sin hogar, escucharon sus historias y entendieron sus necesidades.

1.3.6. Perú

El Gobierno de Perú desarrolló el Modelo de Datos Abiertos Gubernamentales[27] en el marco del modelo de gestión para resultados orientado al servicio del ciudadano que promueve la Política Nacional de Modernización de la Gestión Pública de 2021[28]. Este es un modelo diseñado considerando un enfoque de procesos y cadena de valor de los datos abiertos, con el propósito de tener una visión completa y compartida de todos los procesos de nivel estratégico, operativo, apoyo o soporte, incluyendo los procesos de medición del desempeño e impacto.

Este modelo tiene como usuarios todas aquellas personas que tienen la capacidad de transformar los datos abiertos gubernamentales en nuevos productos y servicios públicos. Estos datos pueden ser transformados, combinados o relacionados para facilitar la generación de nuevos recursos de información como vistas, tablas dinámicas, gráficos y mapas de fácil interpretación para los ciudadanos. Incluye un proceso de reutilización que cuenta con la realización de eventos, la creación de espacios y la aplicación de instrumentos o mecanismos que promuevan la reutilización de los datos abiertos (hackathones), el registro, la publicación y la promoción de los productos o los servicios desarrollados a partir de la reutilización de los datos abiertos gubernamentales, reconociendo la autoría del desarrollo y las condiciones de uso.

Para la implementación de los datos abiertos, es necesario formular instrumentos de gestión, alineamiento con las políticas, los planes, las estrategias y las agendas nacionales de la sociedad de la

[25] Disponible en: https://carleton.ca/cuids/cu-events/crowdsourcing-with-statistics-canada-a-pilot-project-to-explore-new-frontiers-in-data-collection-mapping-and-open-data/

[26] Disponible en: https://www.infrastructure.gc.ca/homelessness-sans-abri/reports-rapports/pit-counts-dp-2020-2022-highlights-eng.html

[27] Disponible en: https://www.datosabiertos.gob.pe/modelo.pdf

[28] Disponible en: https://cdn.www.gob.pe/uploads/document/file/353854/PNMGP.pdf

información y la competitividad. Además, se deben considerar los lineamientos, los planes y las normas técnicas que emita la Presidencia del Consejo de Ministros, como ente rector.

A continuación, se relacionan una serie de características que deben ser tenidas en cuenta para la aplicación y la optimización de los datos abiertos:

- Infraestructura tecnológica: es importante contar con una plataforma que facilite la publicación, el acceso y la reutilización de los datos abiertos.

- Desarrollo de capacidades y asistencia técnica sobre datos abiertos: esto incluye el diagnóstico de necesidades, los eventos y la evaluación de las capacitaciones y la asistencia técnica dirigida a los servidores civiles involucrados en la apertura y la utilización de datos abiertos de las unidades de tecnologías de la información.

- Promoción y difusión de los datos abiertos, desarrollo de estrategias comunicacionales, de redes sociales y otros canales digitales, eventos y concursos, campañas de difusión en el sector académico, participación en programas de capacitación dirigidos a emprendedores, etc.

- Monitoreo y evaluación, el ente rector debe desarrollar un sistema de monitoreo y evaluación para evaluar los avances de la implementación en función de los indicadores de desempeño: indicadores de impacto (misión y visión), indicadores de resultado (objetivos generales y específicos) e indicadores de producto (acciones). Adicionalmente debe establecer mecanismos de supervisión y control de calidad de los datos.

- Identificación de necesidades de los datos, identificar los usuarios de los datos, comunidades de desarrolladores, comunidad académica, empresas, organizaciones de la sociedad civil, etc.

- Recopilación, incluye actividades de identificación de los conjuntos de datos disponibles, en análisis y priorización de los datos y la elaboración del catálogo de datos abiertos conformado por los conjuntos de datos reutilizables, organizado por categorías o temáticas y periodos de actualización como parte de los metadatos.

- Proceso de tratamiento, comprende la revisión del catálogo de datos y su control de calidad, incluye la contextualización, la interpretación, la depuración o la conversión de los conjuntos de datos a un formato abierto antes de ser publicado.

- Los conjuntos de datos deben ser actualizados periódicamente para asegurar su permanencia para mantener el interés de las comunidades de usuarios.

- Para la gestión y la sostenibilidad del proceso de apertura y reutilización de datos abiertos es necesario habilitar canales de gestión de cambio para fortalecer la cultura basada en datos, en el conocimiento abierto y centrada en el ciudadano, buscando mejorar su calidad de vida y brindando oportunidades de desarrollo.

1.3.7. OCDE

La Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha publicado varias guías sobre el uso de DGC para la elaboración de políticas públicas. Estas guías se basan en la idea de que los DGC pueden ser una fuente valiosa de información para los gobiernos, ya que pueden proporcionar información sobre las necesidades y las preferencias de los ciudadanos de una manera más rápida y económica que los métodos tradicionales de recopilación de datos.

Las guías de la OCDE sobre DGC cubren una amplia gama de temas, incluyendo:

- ¿Cómo recopilar y utilizar DGC de manera responsable?
- ¿Cómo garantizar la calidad y la confiabilidad de los DGC?
- ¿Cómo proteger la privacidad de los ciudadanos?
- ¿Cómo utilizar DGC para mejorar la transparencia y la rendición de cuentas del gobierno?

A continuación, se presenta un resumen de algunas de las guías más importantes de la OCDE sobre DGC:

Guía de la OCDE sobre la Gobernanza de Datos Generados por la Ciudadanía (2019): esta guía proporciona un marco para que los gobiernos puedan desarrollar políticas y prácticas para la gestión responsable de DGC.[29]

Guía de la OCDE sobre la Calidad y la Confiabilidad de los Datos Generados por la Ciudadanía (2020): esta guía ofrece recomendaciones para que los gobiernos puedan garantizar que los DGC que utilizan sean de alta calidad y confiables.[30]

Guía de la OCDE sobre la Privacidad y los Datos Generados por la Ciudadanía (2021): esta guía proporciona orientación a los gobiernos sobre cómo proteger la privacidad de los ciudadanos al recopilar y utilizar DGC. [31]

[29] Disponible en: www.oecd.org/gov/dominican-republic-scan.pdf

[30] Disponible en: https://www.oecd-ilibrary.org/sites/f1b22902-es/index.html?itemId=/content/publication/f1b22902-es

[31] Disponible en: https://www.oecd.org/tax/transparency/documents/confidentiality-ism-toolkit_es.pdf

Guía de la OCDE sobre la Transparencia y la Rendición de Cuentas del Gobierno a través de los Datos Generados por la Ciudadanía (2022): esta guía describe cómo los DGC pueden utilizarse para mejorar la transparencia y la rendición de cuentas del gobierno.[32]

Las guías de la OCDE sobre DGC son un recurso valioso para los gobiernos que buscan aprovechar el potencial de esta nueva fuente de datos. Las guías proporcionan información práctica y recomendaciones sobre cómo recopilar, utilizar y proteger DGC de manera responsable.

La OCDE reconoce la importancia de los DGC para mejorar la toma de decisiones, la transparencia y la participación pública. En este contexto, la OCDE ha desarrollado diversas iniciativas para:

1. Facilitar la generación y el acceso a los DGC:

- Marco de Datos Abiertos: la OCDE promueve la adopción de principios de datos abiertos para que los datos públicos sean accesibles, utilizables y comparables.
- Plataformas de datos: la OCDE ofrece plataformas como data.oecd.org que permiten el acceso y la visualización de datos de diversos temas, incluyendo algunos DGC.
- Herramientas y guías: la OCDE proporciona herramientas y guías para ayudar a los ciudadanos a generar, compartir y utilizar DGC.

2. Fomentar el uso de los DGC para la innovación:
    - Concursos y desafíos: la OCDE organiza concursos y desafíos que invitan a los ciudadanos a utilizar DGC para desarrollar soluciones innovadoras a problemas sociales.
    - Financiamiento: la OCDE ofrece financiamiento para proyectos que utilizan DGC para mejorar la transparencia, la rendición de cuentas y la participación pública.

3. Abordar los desafíos relacionados con los DGC:

- Privacidad y seguridad: la OCDE reconoce la importancia de proteger la privacidad y la seguridad de los DGC.
- Calidad de los datos: la OCDE trabaja para mejorar la calidad y la comparabilidad de los DGC.
- Capacidades digitales: la OCDE promueve el desarrollo de capacidades digitales para que los ciudadanos puedan generar, utilizar y comprender los DGC.

Ejemplos de DGC en la OCDE:

- Datos de sensores ambientales: los ciudadanos pueden utilizar sensores para recopilar datos sobre la calidad del aire, el ruido y otros indicadores ambientales.
- Datos de transporte: los ciudadanos pueden compartir datos sobre sus viajes para mejorar la planificación del transporte y la gestión del tráfico.

[32] Disponible en: https://www.oecd.org/tax/transparency/documents/informe-anual-foro-global-2022.pdf

- Datos de salud: los ciudadanos pueden compartir datos sobre su salud y bienestar para mejorar la investigación médica y la prestación de servicios de salud.

Las Directrices sobre Participación Ciudadana de la OCDE están dirigidas a cualquier funcionario o institución pública con intención de llevar a cabo un proceso de participación ciudadana. Las directrices describen diez pasos para diseñar, planificar, implementar y evaluar un proceso de participación ciudadana, y describen ocho métodos diferentes para implicar a los ciudadanos: información y datos, cabildos abiertos, consultas públicas, innovación abierta, ciencia ciudadana, monitoreo cívico, presupuestos participativos y procesos deliberativos representativos. Las directrices se ilustran con ejemplos y orientaciones prácticas basadas en datos recopilados por la OCDE. Por último, se presentan nueve principios rectores para ayudar a garantizar la calidad de estos procesos.[33].

1.4. Conclusiones

- Las estrategias presentadas por Uganda para la transformación de datos generados por la ciudadanía en estadísticas oficiales, considerando la calidad de los datos, son un importante referente para otras naciones que esperan incluir información procedente de la sociedad civil de manera adecuada en la información oficial y se resalta el enfoque de género e igualdad que reconoce el marco normativo de esta nación.
- En el contexto de los portales de datos abiertos en Europa, algunos conjuntos de datos DGC están disponibles, pero los datos generados por proyectos de ciencia ciudadana (CS) rara vez se incluyen[34].
- Muchos portales de datos abiertos incluyen conjuntos de datos que se generan como resultado de iniciativas de participación ciudadana (encuestas, quejas, etc.). Aunque estos conjuntos de datos contienen DGC, estas iniciativas suelen ser iniciadas por las administraciones y los ciudadanos quedan relegados a meros proveedores[35].
- Existen oportunidades interesantes para incluir DGC en los portales de datos abiertos, en general, y en data.europa.eu en particular. Siguiendo estas recomendaciones, la cantidad de DGC no sólo aumentaría, sino que también promovería potencialmente la participación ciudadana.[36]
- La guía sobre datos generados por ciudadanos en Kenia, respaldada por la Oficina Nacional de Estadísticas (KNBS) y potenciada por la participación de las organizaciones de la sociedad civil, establece un marco integral para la generación y la utilización de datos de calidad. La KNBS

[33] Disponible en: https://www.oecd.org/publications/directrices-de-la-ocde-sobre-procesos-de-participacion-ciudadana-f1b22902-es.htm

[34] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

[35] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

[36] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

desempeña un papel esencial al proporcionar asesoramiento técnico, marcos de muestreo, y garantizar la calidad de los datos generados. Mientras tanto, las organizaciones de la sociedad civil contribuyen al proceso mediante la facilitación de la participación ciudadana, el desarrollo de capacidades y la colaboración activa. Los beneficios derivados de la incorporación de datos generados por ciudadanos en los procesos de toma de decisiones y políticas incluyen la inclusividad, la adaptación de políticas a necesidades reales, mayor transparencia y rendición de cuentas, destacando la importancia de esta colaboración multisectorial para un enfoque más informado y centrado en la comunidad en el ámbito de la toma de decisiones.

- Los DGC ofrecen un enorme potencial para mejorar la vida de los ciudadanos, pero su uso efectivo requiere superar desafíos como la falta de marco legal, capacidad técnica y confianza. El gobierno debe actuar para crear un entorno propicio que fomente la innovación y el desarrollo con DGC, teniendo siempre como prioridad la privacidad y la seguridad de los datos.
- Aunque no se encontró una documentación formal que de lineamientos sobre el uso de DGC para su aprovechamiento estadístico por parte de Statistics Canada, se evidenció que esta oficina reconoce su importancia, y los utiliza para en proyectos como la medición de la canasta de mercado, las encuestas de impacto del COVID-19, el Open Database of Buildings y Everyone Counts.

1.5. Recomendaciones

Se recomienda en los procesos de generación de guías, documentos y marcos normativos con relación a datos generados por la ciudadanía:

- Considerar un enfoque diferencial de género, dado que este facilita la inclusión de todas las personas en las estadísticas oficiales de las naciones.
- Buscar activamente activos valiosos de DGC a través de convocatorias abiertas y asociaciones con actores clave de la informática. como ECSA y oficinas nacionales y regionales en CS y proyectos de CS[37].
- Facilitar el descubrimiento de DGC en portales de datos abiertos etiquetando todos los conjuntos de datos de DGC con una etiqueta específica como 'DGC' o 'datos generados por ciudadanos'[38].
- Establecer procedimientos para capturar procesos de DGC y métodos de validación de datos para aumentar la confianza de los usuarios de datos de terceros[39].

[37]Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

[38] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

[39] Disponible en: https://data.europa.eu/sites/default/files/report/data.europa.eu_Report_Citizen-generateddataondata_europa_eu.pdf

- Desarrollar un marco legal y ético sólido que regule la recolección, el uso y el almacenamiento de datos generados por los ciudadanos. Este marco debe garantizar la privacidad, la seguridad y la no discriminación en el uso de los datos.
- Establecer mecanismos de transparencia y rendición de cuentas que permitan a los ciudadanos conocer cómo se están utilizando sus datos y cómo pueden ejercer sus derechos sobre ellos.
- Promover la educación y la formación en materia de datos abiertos para que la ciudadanía pueda comprender los beneficios y los riesgos asociados a su uso.
- Establecer estándares de interoperabilidad para que los datos generados por diferentes entidades públicas y privadas puedan ser fácilmente compartidos y utilizados.
- Revisar si las fuentes de información consideradas como fuentes de DGC (*crowdsourcing*, datos abiertos y encuestas generadas en articulación con la ciudadanía) son compatibles con el marco de Copenhague, y si no es así, considerar si son fuentes aplicables para el contexto colombiano.

Revisión de
REFERENTES
INTERNACIONALES

2.

Metodología para calcular el indicador ODS 10.7.3: número de personas que murieron o desaparecieron en el proceso de migración hacia un destino internacional

2. Reseña: metodología para calcular el indicador ODS 10.7.3: número de personas que murieron o desaparecieron en el proceso de migración hacia un destino internacional

Expositores:

- Eric Manuel Rodríguez Herrera, director de Planeación de la Dirección de Planeación, Junta de Gobierno y Presidencia INEGI.
- Naghielli Angélica Álvarez Chombo, jefa de Departamento de Análisis de Información Estadística y Geográfica, Junta de Gobierno y Presidencia INEGI.

Importancia de la medición del ODS 10.7.3

Este indicador busca visibilizar y caracterizar el fenómeno de las personas migrantes que mueren o desaparecen mientras transitan por México con la intención de llegar a Estados Unidos. Dada la posición geográfica de México, se ha convertido en un corredor migratorio donde los migrantes en situación irregular están expuestos a riesgos y vulnerabilidades asociadas a la violencia, el crimen organizado, los accidentes de transporte, la deshidratación y demás peligros propios del trayecto.

Por ello, es fundamental dimensionar la magnitud de este fenómeno para diseñar mejores políticas públicas de prevención, atención y asistencia a los migrantes. Asimismo, contar con información confiable permite caracterizar sociodemográficamente a esta población y determinar perfiles de riesgo.

Aspectos metodológicos

Ante la falta de un registro único de migrantes en tránsito en México, la medición del INEGI se basa principalmente en el registro de defunciones que ya recopila esta entidad. Esta fuente de datos permite tener una medida año a año de la mortalidad en el país.

El registro de defunciones se complementa con otras fuentes como el Censo de Población y Vivienda, encuestas específicas de migración en fronteras y datos de detección de migrantes en situación irregular recolectados por la Unidad de Política Migratoria.

Para acotar la población objetivo específicamente a migrantes en tránsito, el INEGI aplica varios filtros a las defunciones registradas:

- Nacionalidad no mexicana: se identifican defunciones de personas extranjeras según su país de nacimiento y nacionalidad.
- No residentes en México: se descartan extranjeros que sí tenían residencia habitual en México para no contabilizar migración permanente.

- Nacionalidades con presencia irregular: usando datos de la Unidad de Política Migratoria se enfocan en nacionalidades que tienen flujos irregulares detectados dentro del país.
- Muertes en municipios de alta presencia de tránsito migratorio: se clasifican municipios según convergencia de rutas migratorias, detecciones de irregulares, accidentes previos, entre otros, para filtrar defunciones que ocurrieron en ellos.

Luego de obtener la población migrante en tránsito, el INEGI analiza variables sociodemográficas como edad, sexo y causa de muerte para caracterizar a esta población y comparar sus patrones con migrantes residentes en México y mexicanos en las defunciones registradas.

Algunas consideraciones adicionales

- Se enfocan por el momento en defunciones y no en desapariciones debido a las implicaciones legales e institucionales diferentes de ambos conceptos.
- Toman en cuenta muertes por causas externas y naturales para no excluir enfermedades previas o efectos del tránsito irregular que pueden producir problemas cardíacos, deshidratación, entre otros.
- Solo analizan migrantes en tránsito no mexicanos hacia Estados Unidos y no incluyen todavía casos de trata de personas. Se podría avanzar en estas desagregaciones en el futuro.

Principales retos

El INEGI ha enfrentado varios retos durante el desarrollo metodológico de la medición del ODS 10.7.3 en México.

- Como se mencionó, no existe un registro único que permita identificar directamente a la población de migrantes en tránsito para hacer su seguimiento.
- Ante los vacíos de información ha sido necesario usar diferentes estrategias para aproximarse a este subregistro como la combinación de múltiples fuentes de datos existentes, la construcción de proxy variables, el uso de supuestos y los criterios de clasificación, así como el apoyo de otras disciplinas como la demografía y la geografía para explotar mejor las fuentes.
- Persiste incertidumbre sobre qué fracción de casos podrían aún estar sobreestimando el fenómeno (por ejemplo, inclusión de algunos turistas extranjeros fallecidos) o subestimándolo al no captar todas las rutas y los municipios de relevancia. Se necesita seguir validando los resultados.
- Falta de claridad y estandarización de algunos conceptos claves como las diferencias entre migrantes en tránsito, migrantes económicos, desplazados forzados y la distinción legal de algunas categorías migratorias en México. Esto dificulta delimitar la población de interés.
- Si bien se ha involucrado a múltiples instituciones y agencias de México en esta medición, aún falta consolidar espacios formales de reporte, publicación, retroalimentación y gobernanza de las cifras.

Vale la pena resaltar algunos aprendizajes en cuanto a buenas prácticas para hacer frente a estos retos:

- Aprovechar fuentes de datos existentes a través del desarrollo de técnicas innovadoras de procesamiento, cruce, filtrado y modelamiento.
- Involucrar instituciones técnicas y académicas especializadas desde el inicio para retroalimentar la metodología.
- Evaluar continuamente los supuestos y las limitaciones en la identificación de la población objetivo.
- Contrastar resultados con otras mediciones e investigaciones afines.
- Documentar y transparentar en detalle los criterios utilizados, sus debilidades y el efecto sobre los datos.

Algunos resultados destacados

Luego de aplicar toda la metodología para 2019, el INEGI encontró en este ejercicio exploratorio varios patrones en la población de migrantes en tránsito fallecidos:

- Predominio de muertes de hombres de un 81,2% y de estos una mayor representación en edad productiva entre los 20 y los 39 años. Ante la creciente participación de mujeres en flujos mixtos recientes, se requieren mayores investigaciones con enfoque de género sobre esta tendencia.
- El 76,7% de las muertes correspondieron a causas externas, principalmente agresiones (32%), accidentes de transporte y ahogamientos. Las causas naturales (infartos) representaron el 23% de los casos.
- Los principales países de origen fueron Honduras, Guatemala, El Salvador y Colombia. Se necesita más investigación cualitativa para entender la presencia de migrantes extrarregionales.
- Un 25% de casos tenía país de nacimiento no identificado, probablemente por indocumentación durante el trayecto. Mejorar la caracterización de estos eventos permitiría comprender mejor el perfil y las vulnerabilidades de esta población invisible para los registros regulares.
- Los estados con más casos fueron Chiapas, Tamaulipas y Quintana Roo. Sin embargo, al separar este último estado con alta presencia turística, se encontraron patrones contrastantes en cuanto a nivel educativo, ocupación y tipo de institución donde se registró la defunción.

La información recopilada hasta ahora permite confirmar varias hipótesis sobre las condiciones de vulnerabilidad que enfrentan las personas migrantes en situación irregular en su trayecto por México, en línea con lo documentado por organizaciones de la sociedad civil.

Sin embargo, los resultados son preliminares en tanto la metodología sigue en desarrollo. Por el momento constituyen un primer acercamiento útil para la generación de nuevas preguntas y la orientación de políticas públicas en la materia.

Importancia para el DANE

Conocer la experiencia del INEGI permite al DANE identificar buenas prácticas, vacíos de información y oportunidades de mejora para fortalecer la medición del indicador ODS 10.7.3 en Colombia.

En primer lugar, resalta la importancia de aprovechar mejor los registros administrativos que ya recolectan las entidades nacionales, más allá de los datos que se reportan a nivel internacional desde instancias como la Organización Internacional para las Migraciones (OIM). El INEGI logró explotar el potencial de fuentes existentes, pero subutilizadas, en especial los registros de defunciones y aclaró sus limitaciones.

También evidencia la utilidad de construir indicadores proxy ante fenómenos como la migración irregular que son de difícil captación por su naturaleza escasa y vulnerable. Mediante la triangulación creativa de datos, el diseño de algoritmos de clasificación y la aplicación de criterios sociodemográficos y geoespaciales, el INEGI se aproxima al subregistro.

Otro aprendizaje fundamental es la configuración de equipos interdisciplinarios (demógrafos, geógrafos, epidemiólogos, *data scientists*, etc.) para enriquecer los análisis cuantitativos con perspectivas coyunturales, mapeos detallados, análisis de series de tiempo y revisiones de la calidad del dato para detectar tempranamente sesgos y oportunidades.

Asimismo, la comunicación cercana con pares internacionales como la OIM permite compatibilizar conceptos, estandarizar variables e identificar sinergias entre distintas mediciones globales, regionales y nacionales del fenómeno migratorio.

Por último, a nivel institucional, conformar mesas de trabajo interinstitucionales para la gobernanza de este indicador podría garantizar la sostenibilidad, la trazabilidad y la mejora gradual de la medición. Tal instancia facilitaría el reporte y potenciaría la capacidad técnica institucional de todas las entidades participantes.

Esta presentación no solo permite al DANE reconocer vacíos de información concretos y opciones viables para la adaptación de esta metodología al contexto colombiano. También brinda luces en términos del proceso colaborativo requerido entre productores y usuarios de datos, para mejorar la actual captación estadística de un fenómeno invisible y decisivo dentro de la movilidad humana regional e internacional. El DANE valora enormemente esta cooperación técnica con el INEGI y su buena disposición para compartir su experticia y lecciones aprendidas en la conceptualización y la implementación de esta medición innovadora en México. Los hallazgos presentados motivan al equipo colombiano a iniciar un trabajo interinstitucional para desarrollar una propuesta propia que aproveche al máximo los registros administrativos existentes y logre visibilizar a los migrantes que pierden la vida durante su tránsito por el país.

Revisión de
REFERENTES
INTERNACIONALES

3.

Reseña: desafíos en la producción de estadísticas oficiales con nuevos métodos de recolección de datos

3. Reseña: desafíos en la producción de estadísticas oficiales con nuevos métodos de recolección de datos[40]

Expositores:

- Profesor Distinguido de Investigación Danny Pfefferman, Profesor de Estadística Social, Universidad de Southampton, Reino Unido y Profesor Emérito, Departamento de Estadística y Ciencia de Datos, Universidad Hebrea de Jerusalén, Israel.
- Profesor Distinguido de Investigación J N K Rao, Escuela de Matemáticas y Estadística, Universidad de Carleton, Ottawa, Canadá (Moderador).

Introducción

El seminario inicia con la presentación del profesor Danny Pfefferman describiendo brevemente su experiencia como estadístico y director durante 9 años de la Oficina Central de Estadísticas de Israel. En ese rol debió enfrentar retos como la pandemia de COVID-19 que obligó a reducir el equipo de trabajo presencial, implementar el trabajo remoto y adaptar las encuestas para recolectar datos críticos para el gobierno de forma oportuna.

El profesor Pfefferman abordó conceptos clave relacionados con las estadísticas oficiales que se basan en datos recolectados por las oficinas nacionales de estadística a través de encuestas probabilísticas o registros administrativos principalmente. Estas estadísticas son esenciales para la formulación de políticas públicas y la toma de decisiones. Sin embargo, en años recientes, los cambios tecnológicos y el surgimiento de nuevas fuentes de datos representan desafíos importantes para producir estadísticas confiables, oportunas y relevantes.

Durante el seminario, se discutió sobre los métodos tradicionales de recolección de datos para estadísticas oficiales, como las encuestas por muestreo probabilístico y los registros administrativos. Además, se mencionó cómo las oficinas de estadística tuvieron que adaptarse durante la pandemia de COVID-19, implementando nuevas metodologías y procedimientos para recopilar datos, como encuestas telefónicas y en línea. También se abordó la integración de múltiples registros administrativos para la producción de estadísticas oficiales, resaltando los desafíos y la importancia de proteger la privacidad de los datos. Estos temas destacan la relevancia de las estadísticas oficiales, los desafíos en su producción y la necesidad de adaptarse a cambios tecnológicos y fuentes de datos emergentes.

[40]Disponible en: https://sce.org.co/challenges-in-the-production-of-official-statistics-with-new-methods-of-data-collection-a-talk-by-prof-danny-pfeffermann/

Principales retos de las oficinas de estadística nacional

Uno de los principales retos mencionados fue obtener estimadores insesgados a partir de muestras probabilísticas que permiten calcular medidas de precisión sin necesidad de supuestos estadísticos. No obstante, a menudo se requieren muestras de gran tamaño para estimaciones confiables a nivel de subpoblaciones, lo cual tiene un alto costo.

Otro reto se relaciona con posibles sesgos en la selección de las muestras o en las respuestas de los encuestados. Por ejemplo, quienes acceden a participar en una encuesta podrían tener características sistemáticamente distintas a quienes se rehúsan, lo que sesgaría las estimaciones.

Igualmente, la creciente renuencia de los hogares y las empresas a participar en encuestas probabilísticas representa un desafío importante. Las tasas de respuesta, antes cercanas al 90%, ahora han descendido al 20-30% en algunos casos.

Por otro lado, las encuestas proxy, en las cuales otra persona del hogar responde por la persona seleccionada en la muestra, conllevan sesgos potenciales. Un estudio presentado mostró diferencias importantes en estimaciones de empleo y desempleo entre respuestas propias y respuestas proxy.

Otro reto mencionado fue la demanda creciente de datos oportunos frente a los tiempos prolongados que toman las encuestas tradicionales desde la recolección hasta la publicación de resultados. Por ejemplo, datos de ingresos de empresas provenientes de registros administrativos pueden obtenerse con sólo dos años de retraso, mientras que algunas encuestas toman más de un año en producir estadísticas.

Nuevos métodos

Respecto a nuevos métodos, se mencionó el uso de registros administrativos y de Big Data. Los registros administrativos proveen datos oportunos y con buena cobertura, pero presentan desafíos como la necesidad de integrar varias bases de datos, lo cual es complejo, así como inquietudes relacionadas con la privacidad de los datos.

En cuanto al Big Data, si bien tiene un gran potencial, aún no se ha implementado en la producción rutinaria de estadísticas oficiales. Entre los inconvenientes están los sesgos de cobertura y selección, su naturaleza no estructurada, la dificultad para acceder a ciertas fuentes de Big Data, y los riesgos de manipulación o divulgación indebida de datos.

No obstante, se recomendó considerar la combinación del Big Data con muestreos probabilísticos tradicionales, así como técnicas para lidiar con la no representatividad de fuentes no probabilísticas, como el emparejamiento de datos o modelos de probabilidad de inclusión.

Resultados y recomendaciones

En cuanto a resultados concretos, el profesor Pfefferman presentó un método bayesiano[41] novedoso para modelar encuestas proxy y encuestas con modos mixtos de recolección (internet, teléfono, presencial). Este enfoque permite obtener estimadores insesgados al tiempo que se modelan los efectos de medición y selección.

De igual manera, propuso un procedimiento para analizar muestras no probabilísticas sin necesidad de supuestos fuertes de ignorabilidad, mediante el modelado de la probabilidad de inclusión dado las variables de interés. Si bien este enfoque luce promisorio, se requiere más investigación teórica y aplicada para evaluar su validez.[42]

Por último, planteó la necesidad de considerar técnicas de *machine learning*, tanto para análisis de series de tiempo como para la producción de estadísticas oficiales en general.

Importancia para el DANE

Este webinar permitió al DANE y a la comunidad estadística conocer los principales desafíos que enfrentan actualmente los institutos nacionales de estadística en varios países para producir datos confiables, ante los cambios en las fuentes de información disponibles. También, se presentaron enfoques innovadores que buscan aprovechar estas nuevas fuentes de datos al tiempo que se preserva la calidad y la rigurosidad de las mediciones.

El contenido resulta de mucha utilidad para el DANE dado sus esfuerzos por modernizar sus operaciones estadísticas, implementar nuevas tecnologías y aprovechar fuentes de información alternativas a las encuestas tradicionales.

Referentes por temática

A continuación, se presentan los referentes abordados en el webinar, organizados por temática, donde se exploran análisis estadísticos clave. En el ámbito de registros administrativos, el trabajo de De Leeuw (2018) destaca los desafíos y las oportunidades de la recolección de datos mediante modos mixtos. Además, Rivers (2007) aborda la integración de registros administrativos y las dificultades asociadas. En relación con el uso de Big Data para estadísticas oficiales, Kim y Wang (2019) examinan técnicas de muestreo. Los métodos para abordar la no representatividad en muestras no probabilísticas son discutidos por Kim y Morikawa (2023). En el contexto de la inferencia desde muestras no probabilísticas,

[41] Disponible en: Pfeffermann, D., & Preminger, A. (2021). Estimation Under Mode Effects and Proxy Surveys, Accounting for Non-ignorable Nonresponse. Sankhyā: The Indian Journal of Statistics, 83-A(2), 779-813.

[42] Disponible en: Kim, J. K., & Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. arXiv:2211.02998v2 [stat.ME]. y Pfeffermann, D., & Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. Sankhya, Series B, 61, 166–186.

Pfeffermann y Sverchkov (1999) proponen enfoques paramétricos y semiparamétricos. La conclusión general destaca la importancia de estos referentes en el abordaje de desafíos estadísticos, respaldada por la obra de Pfeffermann (2015), Sikov (2011) y Rao (2021), quienes profundizan en cuestiones metodológicas y desafíos en la producción de estadísticas oficiales.

Registros administrativos:

- De Leeuw, E.D. (2018). Mixed mode: past, present, and future. Survey Research Methods, 12, 75-89.

Integración de registros administrativos:

- Rivers, D. (2007). Sampling for web surveys. In ASA Proceedings of the Section on Survey Research Methods. American Statistical Association. Alexandria, VA, pp. 4127-4134.

Uso de Big Data para estadísticas oficiales:

- Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis. International Statistical Review, 87, 177-191.

Métodos para abordar la no representatividad:

- Kim, J. K. and Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. Calcutta Statistical Association Bulletin, 75. (To appear.)
- Rao, J.N.K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. Sankhya, Series B, 83, 242-272.
- Pfeffermann, D. and Sverchkov M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. Sankhya, Series 8, 61, 166-186.

Efectos del modo de captura:

- Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. The Journal of Survey Statistics and Methodology (JSSAM), 3, 425-483.
- Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under non-ignorable nonresponse in household surveys with missing covariate information. Journal of Official Statistics. 27, 181-209.
- Pfeffermann, D. and Preminger. A. (2021). Estimation under Mode Effects and Proxy Surveys, Accounting for Non-ignorable Nonresponse. Sankhya, Series A, 83, 779-813.

Revisión de
REFERENTES
INTERNACIONALES

4. Diez propuestas sobre el aprendizaje automático en la estadística oficial

4. Diez propuestas de aprendizaje automático en las estadísticas oficiales[43]

Arnout Van Delden, Joep Burger y Marco Puts

Las discusiones más recientes respecto a los aspectos de calidad relacionados con el uso de Machine Learning (ML) en las estadísticas oficiales se han enfocado en sus implicaciones para los marcos de calidad existentes. Los autores mencionan estar en favor de utilizar ML en las estadísticas oficiales, aunque el interrogante principal continúa siendo qué factores se deben tener en cuenta al utilizar este tipo de modelos. Este documento sugiere diez proposiciones en relación con el uso de ML en las estadísticas oficiales con el propósito de generar sensibilización sobre el mismo. Igualmente, los autores esperan que estas proposiciones generen concientización, estimulen el debate y mejoren el entendimiento entre los modelos estadísticos y los modelos con ML.

Proposiciones

Facilidad en la explicación

1. **La correlación no es suficiente**

Trabajar con ML tiene dos ventajas principales sobre los modelos estadísticos: la adaptabilidad y la automatización. Se adapta mejor debido a que no estima coeficientes de relaciones hipotéticas, sino que aprende patrones automáticamente por ensayo y error. Esto ha llevado a la noción de que, si se dispone de la cantidad adecuada de datos, la correlación sea suficiente. Sin embargo, los modelos ML entrenados en datos observacionales aun encuentran correlaciones que pueden o no ser causales. Según esto, las relaciones correlacionales pueden ser suficientes para una única predicción, aunque para la formulación de políticas, se debe realizar un caso de causalidad. Las relaciones correlaciones solamente son útiles como un punto de partida para generar hipótesis e investigación adicional. Los autores concluyen que las correlaciones entre características o variables y etiquetas no son suficientes para generar predicciones confiables para las estadísticas oficiales. Esto, teniendo en cuenta que cuando el número de características es considerablemente grande, pueden generarse correlaciones falsas. Por lo tanto, estas correlaciones solamente ofrecen un punto de partida y en su lugar, los experimentos o los argumentos sustanciales pueden ser utilizados para encontrar relaciones causales.

2. **Una caja negra resulta de un modelo complejo, bien sea estadístico o algorítmico**

[43] Disponible en: https://link.springer.com/epdf/10.1007/s11943-023-00330-0?sharing_token=-42JMZAclrU4zs1VUMS5Ove4RwlQNchNByi7wbcMAY7gzlqcpvdVcnUo_1t45YMp4l_X-Kkl708l9TtzVLN6CZCcuePNIGFU5h6i4XiXqj5TTwXHFWYF1CjBKebVE3GKSwNiFWGVOLnKkulUJfuvtnYiGIHOUx7Ee4IMNqiwESc=

Los autores mencionan que a pesar de que ML es usualmente considerado como una caja negra (en donde se conocen las entradas de datos y los resultados, pero no el proceso), no debe descartarse como un modelo para las estadísticas oficiales. De hecho, argumentan que sí es posible conocer el proceso: es un conjunto anidado de sumas ponderadas activadas. Igualmente, sugieren que los métodos estadísticos comúnmente aceptados, donde las distribuciones de los parámetros son incluso más difíciles de interpretar que, por ejemplo, un árbol de decisión binario. No obstante, es importante ser capaces de entender las predicciones obtenidas mediante ML.

3. **La necesidad de facilidad en la explicación en las estadísticas oficiales no descalifica el ML**

En las estadísticas oficiales es importante entender los resultados obtenidos y esto es más difícil de lograr con modelos ML. Sin embargo, esto no los descalifica como aptos para las estadísticas oficiales. A modo de ejemplo, los algoritmos son herramientas poderosas que pueden ayudar a las estadísticas oficiales a explotar datos de texto e imagen y mejorar las predicciones o estimaciones al mapear relaciones no lineales e interacciones complejas en conjuntos de datos grandes y ricos. Por lo que es necesario tener buenos métodos de explicación, que sean ajustados de acuerdo con las necesidades de los usuarios. Un ejemplo es los valores SHAP (SHapley Additive exPlanations). Se concluye que aún existe la necesidad de métodos precisos, pero menos computacionales.

Uso

4. **Utilizar ML sigue implicando habilidades**

Incluso asumiendo que la IA reemplaza a los trabajadores, es necesario contar con personas con capacidades especializadas y con experiencia en los modelos ML para obtener buenos resultados o predicciones de casos no vistos en la población. Esto, teniendo en cuenta que el número de posibilidades es muy amplio, oscilando desde la preparación de los datos hasta la selección del modelo.

5. **La optimización de un único algoritmo es mejor que comparar muchos con configuraciones defectuosas**

Los autores mencionan que, si se desea lograr un buen desempeño en el modelo, no es suficiente con intentar un amplio conjunto de algoritmos de ML en el mismo conjunto de datos. Consideran que es más efectivo hacer una selección pequeña de algoritmos idóneos, teniendo en cuenta el problema en cuestión y dedicar suficiente tiempo para optimizar los algoritmos sobre los hiperparámetros y las características (variables).

Hacer inferencia

6. **La imputación masiva representa un problema**

Al intentar predecir una variable objetivo con un modelo ML para todos los valores de la población, se debe tener en cuenta que estos valores no pueden ser considerados como observaciones sin errores. Si la intención es publicar el resultado de múltiples variables de interés, de las cuales una es imputada masivamente, entonces el modelo ML debe considerar las relaciones entre estas variables de interés y la variable a predecir por el modelo ML. Si el interés es computar totales de campos de variables objetivo, con base en algunas variables de clasificación, entonces se debe revisar si el desempeño del modelo en cada uno de los campos es de suficiente calidad.

7. Los modelos ML ignoran injustamente el proceso de datos ausentes

Para esta proposición, el enfoque se encuentra en las implicaciones de la forma como se seleccionan las unidades en el conjunto etiquetado y en el conjunto de pruebas sobre la inferencia que se puede realizar del modelo de desempeño en la población objetivo. Si las unidades se extraen de acuerdo con un diseño de muestreo conocido, con probabilidades de inclusión conocidas, entonces estas pueden ser utilizadas para relacionar las unidades en el conjunto etiquetado con aquellas de la población objetivo. Cuando las probabilidades de inclusión son desconocidas, entonces, para un mayor entendimiento del impacto en que los elementos de una población, los autores utilizan los patrones de datos ausencias descritos por Rubin (1976). Estos son MCAR (Missing completely at random - donde la probabilidad de que falte un elemento y su etiqueta no depende ni de las características ni de la etiqueta; MAR (Missing at Random - donde esta probabilidad depende de ciertas variables (características) pero, condicionado a esas características no depende de la etiqueta) y MNAR (Missing not at random - donde esta probabilidad depende de la etiqueta (desconocida) de manera que no puede ser explicada por las características (conocidas).

Se concluye, por tanto, que el grado en que se puede generalizar desde el conjunto de pruebas hasta la población objetivo, en el caso en el que las probabilidades de inclusión sean desconocidas, depende del patrón de datos ausentes: MCAR, MAR, o MNAR. Si solamente se tiene una muestra no probabilística, no se puede tener la certeza de cual patrón de datos ausentes aplica, al menos de que se encuentre una variable de antecedentes que explique dicha ausencia y en tal caso, aplicaría MAR. Algunas veces se pueden utilizar argumentos sustanciales sobre los cuales el patrón sea el más probable. En la situación de MNAR, se podrían necesitar datos complementarios para corregir el sesgo potencial, como una muestra probabilística con la variable de interés.

Configuración de datos

8. Los modelos de ML deben ampliarse para abarcar más configuraciones de datos

A modo de resumen, se identificaron tres temas para aplicar ML en las estadísticas oficiales, donde se requieren avances: el uso de características de alto nivel; la predicción de etiquetas, distintas de las nominales o numéricas, y la predicción de clases raras. Estos aspectos no están limitados a las

estadísticas oficiales. Con respecto a la predicción de clases raras, se identificaron cuatro puntos para investigación adicional: estimar la proporción de la clase rara en la población; estratificar la proporción de la clase rara por subpoblaciones; la clase de interés está confundida con otra subpoblación, y la clase negativa es muy diversa.

Calidad

9. Las métricas de calidad de ML son insuficientes para las estadísticas oficiales

Los autores resumen que en las aplicaciones de ML resulta bastante tentador evaluar el desempeño del modelo seleccionando una de las medidas de calidad conocidas. No obstante, en las estadísticas oficiales, el interés se centra en los agregados, razón por la cual aquellas medidas no representan la alternativa más adecuada. Más allá, consideran que el mejor enfoque es pensar cuidadosamente en el objetivo de estudio y a partir de ahí, seleccionar la medida de calidad.

10. Las dimensiones de calidad existentes son suficientes

En esta proposición, se discute que los aspectos de calidad de ML (interpretabilidad, robustez, estabilidad, precisión y validez del modelo) respecto a las estadísticas oficiales encajan en las dimensiones de calidad de Eurostat (2014) y en su código de prácticas.

Conclusiones

De acuerdo con los autores, el ML está aquí para quedarse en las estadísticas oficiales, debido a que es más flexible en patrones de reconocimiento pues se adapta mejor, tanto al número de instancias como al número de variables, y puede procesar datos de texto e imagen de manera más natural que los modelos estadísticos.

A diferencia de las estadísticas, las cuales se centran en definir el modelo a priori e intentar aproximarse a la verdad fundamental actual tan cerca como sea posible, el ML se enfoca en definir un modelo basado en las correlaciones en los datos, asumiendo que resultará en una presentación precisa de la misma verdad fundamental. Estas diferencias en términos de enfoque han llevado a ciertas discrepancias entre ambos campos.

Es imprescindible resaltar donde se originan las diferencias entre el ML y las estadísticas. El ML ha logrado increíbles avances, sin embargo, las cuestiones relativas a la validez y la utilidad de estos modelos en determinadas circunstancias no han sido fundamentales para el desarrollo de nuevos métodos en este campo.

Este documento contribuye a una mejor utilización del ML. Igualmente, un mejor entendimiento de la selección de algoritmos y afinación ayudará a los estadísticos a utilizarlo mejor.

En la preparación del Reporte de esta edición participamos los siguientes funcionarios:

Diana Marcela Pinzón Topía – dmpinzont@dane.gov.co

Alexander Gonzalez Coca – agonzalezco@dane.gov.co

Catherine Avila Alvarado – jcavilaa@dane.gov.co

Omar Alexander Beltran Vanegas - oabeltranv@dane.gov.co

Gildardo Andres Vargas Acuña - gavargasa@dane.gov.co

Yinneth Mahecha Monsalve - ymahecham@dane.gov.co

Alexandra Jane Simpson Silva - ajsimpsons@dane.gov.co

Julian David Garcia Gomez - jdgarciag@dane.gov.co

Yennifer Dayana Castillo Murcia - ydcastillom@dane.gov.co

Revisión de estilo por: Sonia Naranjo - smnaranjom@dane.gov.co

Revisión de contenido por: Andrea Milena Roncancio Sánchez - amroncancios@dane.gov.co

Si tiene dudas comentarios o aportes sobre esta edición por favor no dude en comunicarse al correo: ymahecham@dane.gov.co - oabeltranv@dane.gov.co

@DANE_Colombia

/DANEColombia

/DANEColombia

@DANEColombia

www.dane.gov.co