
El futuro es de todos
DNP
Departamento
Nacional de Planeación
CAF
BANCO DE DESARROLLO
DE AMÉRICA LATINA
AD
Aprovechamiento
de datos
para la toma
de decisiones
en el sector público

APROVECHAMIENTO DE DATOS PARA LA TOMA DE DECISIONES EN EL SECTOR PÚBLICO

Depósito Legal: DC2021001102
ISBN: 978-980-422-240-5

Editores:
DNP
CAF

Alejandra Botero Barco
Directora General
2021- Actualidad

Luis Alberto Rodríguez Ospino
Director General
2019 - 2021

Daniel Gómez Gaviria
Subdirector General Sectorial

Amparo García Montaña
Subdirectora General Territorial

Diana Patricia Ríos García
Secretaria General

Iván Mauricio Durán Pabón
Director de Desarrollo Digital

Viviana Vanegas Barrero
Subdirectora de Prospectiva Digital

Este estudio ha contado con el apoyo de los asesores:
Eduardo Escobar Gutiérrez
Diana Paola Ramírez Roa
Mariana Quevedo Hernández
Hernán David Insuasti Ceballos
Agustín Jiménez Ospina
Pablo Montenegro Helfer
Juan Sebastián Numpaque Cano
Carlos Andrés Rocha Ruiz
Jairo Andrés Ruiz Saenz

Los autores agradecen los valiosos aportes de:
Liliana Fernández Gómez (DNP)
Ramiro Chaparro (Departamento
Administrativo de la Presidencia de la República)
María Lucila Berniell
(CAF-banco de desarrollo de América Latina)
Enrique Zapata
(CAF-banco de desarrollo de América Latina)

Colaboración externa de Alianza CAOBA

Grupo de Comunicaciones y Relaciones Públicas
Luis Segundo Gámez Daza
Coordinador

Diseño y diagramación:
Good, Comunicación para el Desarrollo Sostenible

El futuro es de todos

DNP
Departamento
Nacional de Planeación

BANCO DE DESARROLLO
DE **AMÉRICA LATINA**

AD

Aprovechamiento **de datos** para la toma de decisiones en el **sector público**

CONTENIDO

FIGURAS

TABLAS

APROVECHAMIENTO DE DATOS PARA LA TOMA DE DECISIONES EN EL SECTOR PÚBLICO DNP Y CAF

La pandemia aceleró la transformación digital y, en su corazón, están los datos. Vivimos en un mundo que se vuelve cada vez más "inteligente" o "*smart*", es decir basado en evidencia, con el crecimiento exponencial de los datos, públicos, privados y personales. A cada momento, las personas, las instituciones públicas y las empresas privadas de las que formamos parte generan más datos que en cualquier otro momento en la historia. Estos datos son, más allá de una representación en una planilla de cálculo, en un json o en una imagen, representaciones de la realidad. Estas representaciones, analizadas de manera correcta nos ayudan a entender mejor nuestra realidad, al mundo que nos rodea y a nuestras interacciones en él. Usadas de manera ética y responsable, nos ayudan a diseñar, implementar y evaluar mejores políticas públicas.

Estamos transitando el camino desde una primera generación de gobernanza de datos que enfatiza la apertura de los datos públicos y la protección de los datos personales, a una segunda generación que enfatiza el valor de los datos para las políticas públicas y de su re-uso para generar valor público; es decir de una concepción del dato como derecho a una concepción del dato como activo. Además de vivir en la época humana en la que más datos se generan, hoy también contamos con la capacidad computacional necesaria y las técnicas algorítmicas adecuadas para hacer sentido de esta gran cantidad de datos. No es de sorprender que las empresas tecnológicas más importantes del mundo estén, en su centro, basadas en los datos y que la economía de los datos se haya convertido en el segmento económico de mayor dinamismo y potencial disruptivo. Crucial en este entendimiento es que hoy los datos son usados, además de como herramientas, como un nuevo espacio en el que las organizaciones tienen que desenvolverse para la creación de estrategias, la toma de decisiones y la ejecución de sus actividades. Sin embargo, la calidad de los datos públicos y de los registros administrativos no es siempre la que podría ser y se necesita intensificar los esfuerzos en esa dirección de mejora.

En este contexto, gobiernos en todo el mundo trabajan para adecuar sus administraciones públicas a esta nueva era, la del dato. Por un lado, los sectores públicos trabajan en la construcción de las infraestructuras de datos necesarias para lograr una interoperabilidad

que permita automatizar procesos e incrementar la eficiencia de las operaciones gubernamentales, sobre la base de un marco ético para el uso responsable de los datos que asegure la privacidad, seguridad e integridad de los datos personales en el marco de derechos digitales. Por el otro, buscan crear las capacidades y equipos para dotar al gobierno con las capacidades analíticas e interpretativas capaces de hacer sentido de estas grandes cantidades de datos y utilizarlas en todo el quehacer del gobierno, desde los servicios al ciudadano hasta la consecución de los Objetivos de Desarrollo Sostenible.

El Gobierno de Colombia ha tomado el liderazgo en la región latinoamericana en temas relativos a innovación digital del Estado basada en una buena gobernanza de los datos. Conjuntamente con CAF -banco de desarrollo de América Latina se ha trabajado en iniciativas como "Manos en la Data" para promover un mayor uso de datos en la toma de decisiones a lo largo de distintas etapas del ciclo de políticas públicas y el fomento del ecosistema govtech para impulsar la colaboración con el sistema emprendedor en la resolución de problemas públicos.

En este contexto, nos complace apoyar al gobierno en la articulación e implementación de sus políticas públicas de inteligencia artificial y de infraestructura nacional de datos. La publicación de este informe sobre el *Aprovechamiento de Datos para la Toma de Decisiones en el Sector Público* se enmarca en esta colaboración. Esperamos que la documentación y análisis de la experiencia del Departamento Nacional de Planeación en la implementación de su estrategia de datos sirva para que otras instituciones en el país y en la región se aprovechen de estos aprendizajes y tomen las mejores prácticas del sector. Estamos convencidos que esto contribuirá a fortalecer el liderazgo de Colombia en la transformación digital de sus instituciones públicas.

Carlos Santiso
Director de Innovación Digital del Estado

CAF -banco de desarollo de América Latina.

APROVECHAMIENTO DE DATOS PARA LA TOMA DE DECISIONES EN EL SECTOR PÚBLICO DNP Y CAF

El ejercicio de coordinar y apoyar la planeación de corto, mediano y largo plazo que permita orientar la definición de políticas públicas en los diversos sectores de la economía y la priorización de recursos de inversión, requiere cada vez más de la disponibilidad de datos de calidad para la toma de decisiones objetivas. Así mismo, en el contexto de la pandemia de la COVID-19, el gobierno ha enfrentado enormes desafíos para satisfacer las demandas de los ciudadanos en términos de prestación de servicios y acceso a la oferta institucional, lo cual ha requerido contar con datos de alta calidad que permitan la focalización adecuada de recursos a población pobre y vulnerable y a empresas afectadas por la crisis.

Dada esta creciente necesidad de mejores datos y de capacidades de analítica, durante los últimos años el Departamento Nacional de Planeación ha dedicado grandes esfuerzos a la consolidación de un marco de política para aumentar el aprovechamiento de datos en el país. Lo anterior, con el objetivo de generar valor social y económico, mejorar la prestación de servicios al ciudadano, apoyar los procesos de formulación de política pública basada en datos y mejorar los niveles de eficiencia de uso de datos.

Por lo tanto, es muy importante para nosotros presentar en esta publicación el contexto general de las políticas públicas que se han liderado desde el Gobierno nacional en esta materia. Y, así mismo, dar a conocer las experiencias y lecciones aprendidas por la Unidad de Científicos de Datos del Departamento Nacional de Planeación en la puesta en marcha de proyectos aplicados de analítica de datos, así como proporcionar una guía metodológica para la formulación de proyectos de analítica, que de claridad y sea de utilidad para todas las entidades del sector público interesadas en hacer de la analítica de datos una herramienta importante en su proceso de diseño de política pública.

Como Directora General del Departamento Nacional de Planeación agradezco la colaboración de CAF -banco de desarrollo de América Latina por apoyar la materialización y consolidación de este documento, que esperamos contribuya especialmente a fortalecer la cultura de datos y la toma de decisiones basada en evidencia en las entidades públicas de Colombia y de América Latina.

Alejandra Botero Barco
Directora General
DNP

RESUMEN EJECUTIVO

LOS DATOS SE HAN CONVERTIDO EN EL ELEMENTO CENTRAL DE LA TRANSFORMACIÓN DIGITAL DE LOS PAÍSES EN TODO EL MUNDO Y EL INSUMO PRINCIPAL PARA LA TOMA DE DECISIONES BASADAS EN EVIDENCIA. LA COYUNTURA PROVOCADA POR LA PANDEMIA DE LA COVID-19 VISIBILIZÓ AÚN MÁS LA IMPORTANCIA DE DISPONER DE DATOS DE CALIDAD Y DE IMPLEMENTAR TÉCNICAS DE ANALÍTICA DE DATOS PARA EL DISEÑO E IMPLEMENTACIÓN DE POLÍTICAS PÚBLICAS EN DISTINTOS ÁMBITOS COMO LA SALUD PÚBLICA, LA EDUCACIÓN, EL TRANSPORTE Y EL MERCADO LABORAL.

Por lo anterior, es indispensable que los gobiernos conozcan la importancia de la analítica de datos para mejorar la toma de decisiones, y, así mismo, dominen las herramientas para fortalecer las capacidades en las entidades públicas para mejorar la gestión de sus datos para aumentar su aprovechamiento.

En el marco de lo anterior, la presente publicación tiene el objetivo de mostrar en qué medida el aprovechamiento de datos se constituye en un aliado relevante para el sector público a fin de diseñar e implementar políticas públicas. Este documento está constituido en tres partes presentadas de la siguiente manera:

- En la primera parte se presenta un contexto de la analítica de datos en el contexto internacional y la apuesta del Gobierno de Colombia para posicionar los datos como activo estratégico para impulsar la transformación digital y mejorar la toma de decisiones. Esto se aborda especialmente desde la descripción del Marco de Política Pública Nacional y la experiencia de la Unidad de Científicos de Datos del DNP.

- En la segunda parte del documento se incluye una guía práctica para las entidades públicas, con el fin de que estas mejoren sus capacidades para la explotación de datos e identifiquen la ruta para la formulación de proyectos de analítica de datos.

- La tercera y última parte, incluye el recuento de los resultados de los proyectos de Manos en la Data-Colombia, una iniciativa impulsada por CAF -banco de desarrollo de América Latina, donde se evidencia el impacto que tiene la analítica de datos para atender problemáticas surgidas como consecuencia de la emergencia sanitaria de la COVID-19.

INTRODUCCIÓN

EL AUMENTO DE LA DIGITALIZACIÓN EN DISTINTOS PROCESOS, LA PROLIFERACIÓN DE DISPOSITIVOS CONECTADOS A INTERNET Y SU USO POR PARTE DE UN GRAN NÚMERO DE PERSONAS HAN PERMITIDO QUE LA GENERACIÓN Y DISPONIBILIDAD DE INFORMACIÓN HAYA CRECIDO EXPONENCIALMENTE EN LOS ÚLTIMOS AÑOS. PARA 2020, MÁS DE 4.500 MILLONES DE PERSONAS EN EL MUNDO USAN INTERNET, 3.800 MILLONES SON USUARIAS DE REDES SOCIALES Y 5.190 MILLONES DE PERSONAS USAN TELÉFONOS MÓVILES.

Con respecto al año 2012, la cifra de penetración mundial a Internet pasó del 30% al 60%, y la de redes sociales subió del 22% al 49% (Datereportal, 2020). En términos de datos, se estima que en un día se están enviando alrededor de 500 millones de mensajes en Twitter, se envían 294 millones de correos electrónicos, 65.000 millones de mensajes en WhatsApp y se realizan 5.000 millones de búsquedas en internet (Foro Económico Mundial, 2019).

Así, el paradigma frente al valor de los datos se ha transformado alrededor del *big data*, es decir, conjuntos de grandes volúmenes de datos que, dada su variedad y rápida generación, ofrecen un gran potencial para identificar nuevas necesidades de la ciudadanía y crear diversas soluciones para suplirlas. De acuerdo con cifras de la Unión Europea, aproximadamente el 90% de los datos disponibles en la actualidad se han generado en los últimos 2 años (Mohamed & Weber, 2020).

En consecuencia, los datos se han convertido en activos estratégicos para la generación de valor social y económico tanto en el sector público como en el privado. Son diversos los aportes del *big data* a la creación de valor para el sector público que, por lo general, se orientan a mejorar la prestación de servicios al ciudadano, acompañar el proceso del ciclo de política pública, mejorar los niveles de eficiencia y la toma de decisiones (McKinsey Global Institute, 2016). En el mismo sentido, la experiencia internacional permite visibilizar la utilidad de la explotación y análisis del *big data* en proyectos de diferentes sectores como el de la salud, el ambiental y el educativo y otros.

El aprovechamiento de grandes cantidades de datos en el sector público trae oportunidades para generar

evidencia en el diseño, la implementación y el seguimiento de los programas, los proyectos y las políticas, pero también importantes desafíos. Entre estos retos se encuentra la necesidad de contar en el país, con las competencias indispensables para el análisis y explotación de datos (DNP, 2018); asimismo, la consolidación de un marco jurídico e institucional que brinde las garantías para el intercambio de datos y la protección de la privacidad de los datos de las personas, el aumento de la cultura basada en datos, al igual que la definición de una infraestructura de datos que permita aumentar su disponibilidad de calidad y su intercambio efectivo entre las entidades.

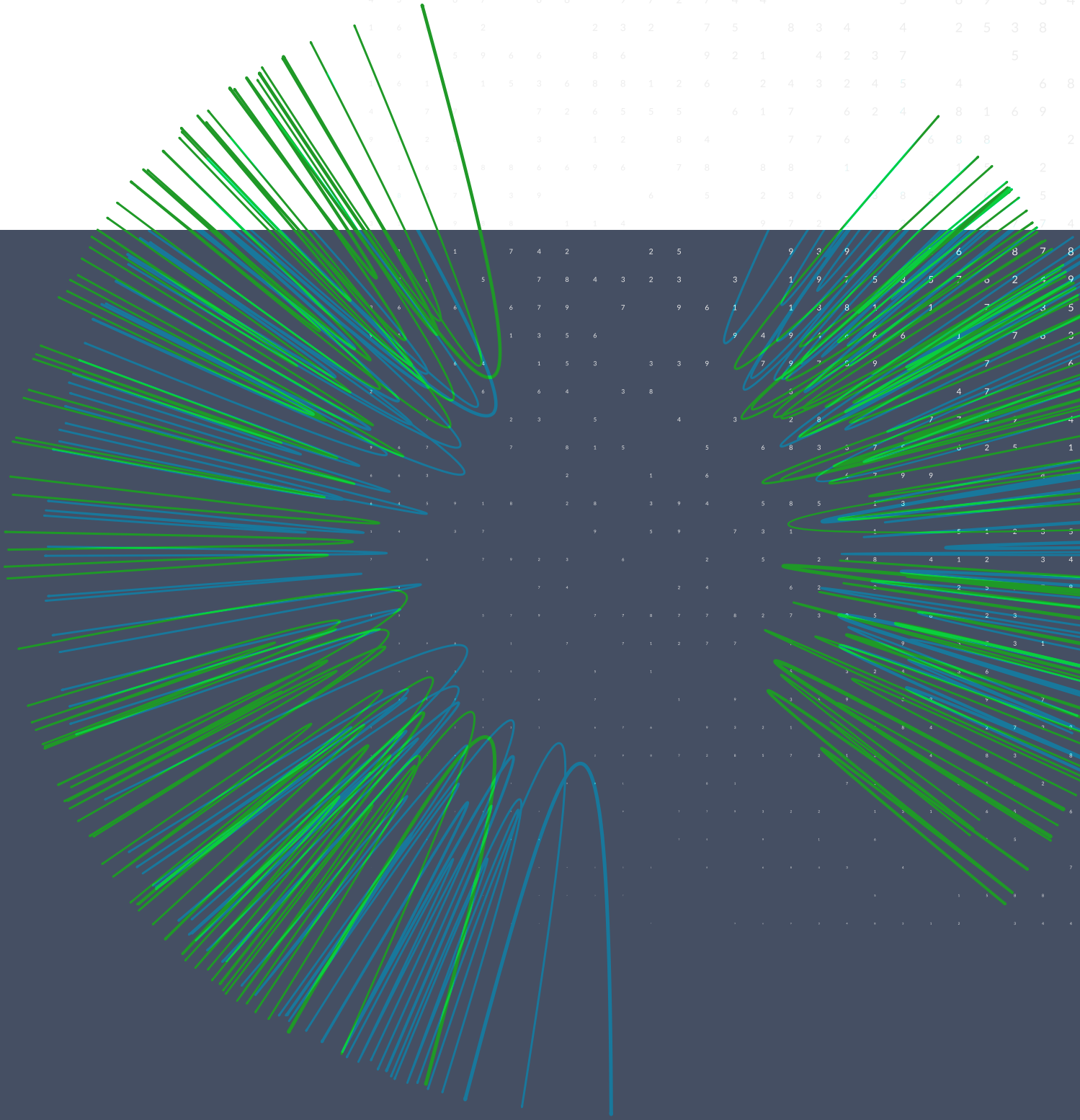
En este contexto, la presente publicación busca brindar a sus lectores un panorama amplio sobre el aprovechamiento del *big data* en Colombia y destacar sus principales aportes para el sector público con base en la experiencia de la Unidad de Científicos de Datos del Departamento Nacional de Planeación. También pretende convertirse en una herramienta práctica para apoyar la implementación de proyectos de explotación de datos en las entidades del sector público del orden nacional y territorial.

El presente documento está compuesto por tres partes:

- **La primera** expone la evolución de la explotación de datos en el contexto global, su aporte para la toma de decisiones en el ciclo de la política pública y las apuestas que se han desarrollado en Colombia para impulsar el aprovechamiento de datos.

- **La segunda** es una guía metodológica para elaborar proyectos de analítica de datos en las entidades públicas del país, con el objetivo de constituirse como una herramienta de apoyo para las entidades que deseen apostarle a la analítica de datos.

- **La tercera** presenta los estudios de caso de la iniciativa *Manos en la Data-Colombia*, financiada por CAF, que muestran la utilidad de la ciencia de datos para la toma de decisiones en política pública, puntualmente en atención de situaciones causadas por la pandemia de la COVID-19.



01

PARTE

Explotación y analítica de datos para el sector público y su evolución en Colombia

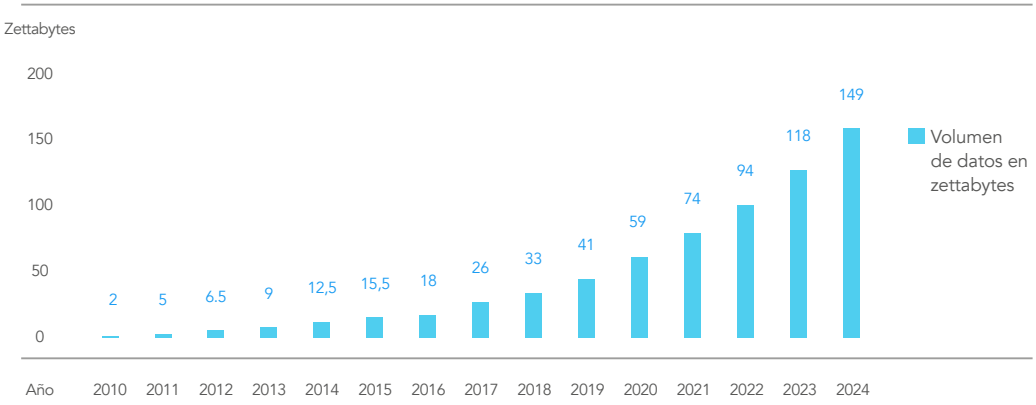
CAPÍTULO 1.1

MARCO CONCEPTUAL DE LA EXPLOTACIÓN Y ANALÍTICA DE DATOS Y SU APLICACIÓN EN EL SECTOR PÚBLICO

CON LA CUARTA REVOLUCIÓN INDUSTRIAL (4RI) LOS DATOS SE HAN CONVERTIDO EN UNO DE LOS ACTIVOS MÁS IMPORTANTES TANTO PARA EL SECTOR PÚBLICO COMO PARA EL SECTOR PRIVADO. EL RÁPIDO DESARROLLO DE LA DIGITALIZACIÓN, LAS NUEVAS TECNOLOGÍAS DIGITALES Y LA ECONOMÍA BASADA EN DATOS SON LOS FACTORES QUE HAN CONTRIBUIDO AL CRECIMIENTO EXPONENCIAL EN LA CREACIÓN Y CONSUMO DE DATOS EN TODO EL MUNDO.

La generación global de datos ha crecido en grandes proporciones durante los últimos 10 años, al pasar de 2 *zettabytes*[1] a 59 *zettabytes* en el año 2020 (figura I.1-1) (Statista, 2020); es más, de acuerdo con las proyecciones del Global DataSphere[2], en el año 2025 se llegará a 175 *zettabytes* creados, capturados y consumidos.

Figura I.1-1. Volumen de datos generados en todo el mundo



Fuente: elaboración propia con base en información de Statista (2020).

El crecimiento exponencial de los datos en todo el mundo obedece más a la copia y consumo que a la creación y generación de datos. De acuerdo con International Data Corporation (IDC), la relación entre la creación y la captura de datos frente a la copia y el consumo es de 1:9, se espera que para el año 2024 esta relación sea de 1:10. Según el análisis del Global

1. Un *zettabyte* (ZB) es equivalente a un trillón de gigabytes.
2. Es un índice que cuantifica y analiza el volumen de datos creados, capturados y replicados a través de distintos años.

DataSphere, la cantidad de datos que se crearán entre el año 2021 y el año 2023 será mayor que los datos creados en los últimos 30 años en el mundo (IDC, 2020).

Por su parte, la firma IDC estima que a finales de 2025 el 80% de los datos mundiales sean datos no estructurados[3] debido, principalmente, al aumento de dispositivos inteligentes de Internet de las cosas (IoT, por su siglas en inglés | IdC por su siglas en español).

La era del *big data* inició su consolidación desde mediados del siglo XX cuando diferentes compañías produjeron tecnologías con el propósito de responder a la necesidad de analizar grandes volúmenes de información. Esos estudios requerían la existencia de sistemas especializados para organizar, sintetizar y transformar información a fin de facilitar los procesos de interpretación de datos. Con base en lo anterior, es posible citar cuatro etapas importantes del siglo XX que permitieron construir las bases para el análisis de *big data:*

1 Aplicaciones independientes - aplicaciones de escritorio,

2 Aplicaciones de escritorio - aplicaciones web,

3 Aplicaciones web - abundancia de aplicaciones de Internet

4 Abundancia de aplicaciones de Internet - aplicaciones de *big data*

(Yaqoob, Targio Hashem, Gani, Mokhtar, & Ahmed, 2016).

La primera fase se desarrolló entre 1960-1990 y se caracterizó por la creación de programas de seguridad de datos, almacenamiento externo rápido y mejoras en la velocidad de transmisión de datos; la segunda, ocurrida desde 1990 a 2000 se caracteriza por el desarrollo de sistemas de recuperación rápida, máquinas o tecnologías para monitoreo y flujo de datos de manera veloz; en la tercera, sucedida de 2000 a 2010, se construyeron sistemas de acceso con características de acceso rápido a datos distribuidos, de uso fácil, con bajos costos y tolerancia a fallos. Por último, en la cuarta fase, comprendida entre 2010 y 2016 se desarrollaron mecanismos de gestión estructurada de datos, accesibilidad rápida a teléfonos inteligentes, comunicación segura, privacidad, visualización, procesamiento en tiempo real, calidad de los datos, comprensión de datos y análisis de datos (Yaqoob, Targio Hashem, Gani, Mokhtar, & Ahmed, 2016). Lo anterior fue necesario para la aplicación de análisis de *big data* y para su desarrollo y consolidación.

En el año 2011, un estudio de Mckinsey Global Institute, titulado *Big data: la próxima frontera para la innovación, la competencia y la productividad*, evidenció el desarrollo y beneficios del *big data* y la analítica de datos, lo cual fomentó un gran interés para el sector privado y los gobiernos, los cuales vieron la oportunidad adelantar dos acciones principales:

1) crear valor social y económico a partir de la gestión de grandes cantidades de datos; y

2) impulsar sus procesos de transformación digital, donde los datos se convierten en activos indispensables para tomar decisiones de manera más rápida e innovar en bienes y servicios. En este contexto, resulta indispensable señalar que es posible aumentar el aprovechamiento de los datos cuando se garantiza su disponibilidad y accesibilidad.

3. No tienen una estructura interna identificable. Los datos no estructurados se pueden almacenar en formatos en PDF, documentos Word, correos electrónicos, datos móviles.

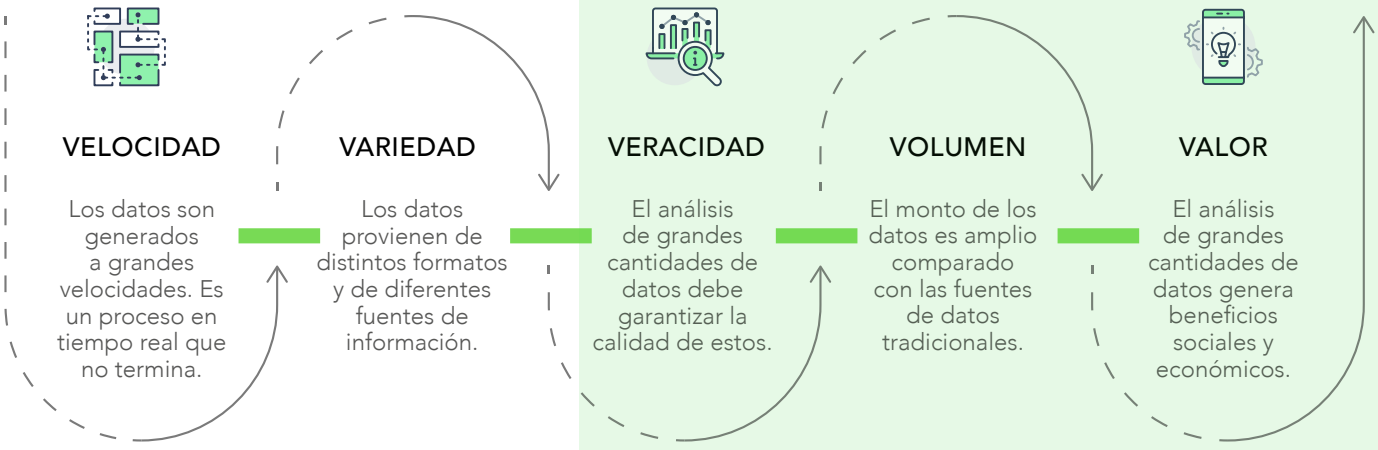
Los gobiernos hacen parte del ecosistema mundial de datos, desde su rol de proveedores de datos a través de la apertura de datos públicos y su rol de consumidores a través de la compilación de datos de los ciudadanos. Ese ecosistema consolidado en los últimos años está integrado, además del sector público, por los proveedores de servicios de internet, los proveedores de infraestructura tecnológica de *software* y *hardware*, los ciudadanos y las empresas privadas —quienes consumen y generan datos— y los proveedores de servicios de analítica de datos. La interacción continua de los actores mencionados aumenta la generación de datos, las posibilidades de reutilización, así como el potencial para su aprovechamiento. A continuación, se hará una breve descripción del *big data*, de sus principales características y de las principales técnicas de análisis de datos.

1.1.1. DEFINICIONES Y CARACTERÍSTICAS DEL *BIG DATA*

En la actualidad hay diferentes aproximaciones que se emplean para describir el concepto de *big data*. McKinsey, Gartner, Techtarget y Teradata (Targio Hassem, Yaqoob, & Badrul Anuar, 2014) describen el *big data* en términos de las características de sus datos y del valor que generan. El volumen, la variedad y la velocidad a la cual se generan y capturan, el potencial para crear valor frente a la innovación tanto de productos como de servicios, y la necesidad de garantizar, más que en cualquier otro contexto, la confiabilidad y veracidad de los datos, son algunas de las características que rodean al *big data* (figura I.1-2).

En otras definiciones, el *big data* también incorpora elementos que trascienden los datos e involucran procesos, técnicas y tecnologías en las organizaciones para el aprovechamiento de los datos. En tal sentido, el *big data* se define, también, como un proceso sociotécnico, en el cual convergen, por una parte, tecnologías y técnicas para abordar los retos de procesamiento y almacenamiento de datos, procesos estructurados para garantizar la implementación de acciones sistemáticas para el aprovechamiento de los datos y, por otra, el recurso humano para ejecutar dichos procesos y a su vez tomar decisiones a partir de los datos. El *big data* está clasificado en categorías que permiten una mejor comprensión de las características descritas en la figura I.1-2. Esta clasificación se relaciona directamente con las fuentes donde se generan los datos y los tipos de datos de acuerdo con su formato.

Figura I.1-2. Características del *big data*



Fuente: elaboración propia con base en Mohamed & Weber (2020).

1.1.1.1. Fuentes de datos

El fenómeno de *big data* se nutre a partir de fuentes de información que se han consolidado en los últimos años gracias a la digitalización, el aumento en la penetración a Internet y el aumento en el uso de dispositivos móviles; igualmente, la interacción en redes sociales, el almacenamiento de datos en la nube, las páginas web y el Internet de las cosas constituyen en la actualidad las principales fuentes de generación de datos masivos.

Ahora bien, la generación de datos puede darse de manera activa, donde el sujeto o los sujetos tienen la intención de producirlos, o una generación pasiva, es decir que los datos no se generan intencionalmente por el usuario, sino que se derivan del uso de dispositivos. A continuación, se mencionan las principales fuentes de generación de datos masivos:

La generación de datos por la interacción humana

A través de las redes sociales y las páginas web, en el que es posible el intercambio de ideas y la creación de comunidades virtuales a través de dispositivos móviles y digitales. Las páginas web generan datos a partir de las búsquedas de información, las transacciones de bienes y servicios, las publicaciones sobre vacantes y oferentes de empleo, y los sitios *on-line* del Gobierno. Estos datos se generan en forma de documentos, fotos, videos, audios y mensajes de texto, entre otros.

Los sistemas de información

Que consolidan datos provenientes de registros administrativos, encuestas nacionales y territoriales y sondeos, entre otras fuentes que generalmente proveen datos de tipo estructurado.

Los datos transaccionales

Que incluyen la información de los precios de las acciones, los datos bancarios, los datos de transacciones financieras, los historiales de compra de los individuos.

Las máquinas y los dispositivos

Que generan datos y los comparten a través de Internet sin intervención humana (IdC). Algunos ejemplos de tales dispositivos son los GPS, los celulares, las tabletas, las cámaras digitales conectadas a Internet, los automóviles, los refrigeradores y las lavadoras inteligentes. En esta categoría también se encuentran los dispositivos de sensores que estiman comportamientos físicos para convertirlos en señales de tráfico, medio ambiente, seguridad, sismología y otros más.

1.1.1.2. Tipos de datos según su formato

Los datos tienen características con respecto a la manera en cómo se presentan, esto respecta datos estructurados, no estructurados y semiestructurados (tabla I.1-1).

Tabla I.1-1. Tipos de datos según su formato

TIPOS DE DATOS	DESCRIPCIÓN	EJEMPLOS
Estructurados	Son datos que tienen una estructura interna identificable. Tienen definida la longitud, el formato y su tamaño.	Hojas de cálculo, bases de datos relacionales y formato tabla, archivos de registro.
No estructurados	No tienen una estructura interna identificable.	Documentos de PDF o Word, mensajes de texto y correo electrónico, grabaciones, videos de audio.
Semiestructurados	No siguen un sistema de bases de datos convencional. Aunque suelen tener la forma de datos estructurados no están organizados en modelos de bases de datos relacionales.	Páginas web, señales de tráfico. Formatos XML y JSON.

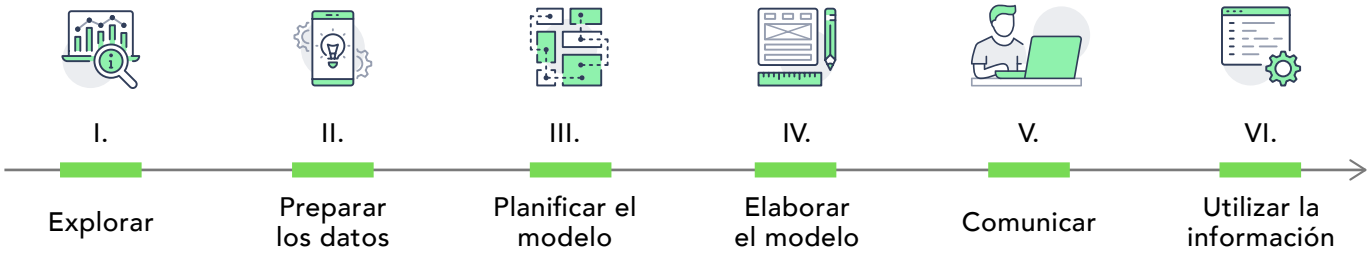
Fuente: elaboración propia.

1.1.1.3. Analítica de datos y *big data*

La generación de grandes volúmenes de datos en distintas fuentes exige formas innovadoras y prácticas para procesar y analizar esa información, de tal manera que pueda comprenderse y tomar decisiones a partir de ella; por tal motivo, las técnicas de análisis de datos se constituyen actualmente en una herramienta indispensable para extraer todo el valor de los datos. La generación de valor a partir de datos se enmarca en el proceso de cadena de valor, el cual comienza con la captura de los datos, pasa por el procesamiento, la comunicación y la utilización de la información. El valor de los datos se obtiene en la etapa de análisis y procesamiento y es allí donde surge la "ciencia de datos", la cual incorpora una serie de técnicas de distintas disciplinas como la minería de datos, la estadística, las matemáticas, el aprendizaje automático y las ciencias de la computación.

La figura I.1-3 representa el ciclo requerido para llevar a cabo un proceso de analítica de datos. Los pasos tienen una relación iterativa, por lo que dentro del esquema se evidencian posibles retrocesos entre las etapas que representan la posibilidad de reformular o transformar el proceso conforme a las características de los datos y de los sistemas (Rodríguez, Palomino, & Mondaca, 2017).

Figura I.1-3. Ciclo de vida del análisis de datos



Fuente: elaboración propia con datos de Rodríguez, Palomino, & Mondaca (2017).

Primera fase

Se exploran los datos, se identifica la problemática que se busca abordar, y se elabora la pregunta y la hipótesis que se pretende responder con los datos.

Segunda fase

Se preparan los datos, se identifica cuáles serán necesarios para abordar el problema, se determina dónde están almacenados esos datos, se efectúa el proceso de limpieza y el análisis de consistencia de los datos.

Tercera fase

Se planifica el modelo, se identifican las variables que ayudarán a predecir o explicar el modelo, se seleccionan los posibles modelos y algoritmos por usar, y se definen las métricas de desempeño del modelo.

Cuarta fase

Se elabora el modelo; esta etapa la integran la prueba y la selección del modelo de acuerdo con el mejor ajuste y significancia, al igual que la validación del modelo.

Quinta fase

En esta etapa del ciclo se comunican los resultados, para ello se desarrollan actividades como la interpretación de los resultados, se generan las visualizaciones adecuadas para comunicar dichos resultados y se hacen recomendaciones de mejora.

Sexta fase

Corresponde a la de utilización de los resultados; en esta última etapa se toman decisiones a partir de la información arrojada por el modelo.

Para el procesamiento de los datos es necesario, en un primer momento, abordar la gestión de datos, que incluye las tareas de su captura, almacenamiento, depuración y organización. El cumplimiento de estas tareas posibilita generar conjuntos de datos con propiedades que faciliten el análisis y su procesamiento.

En un segundo momento, se procede a implementar procesos de analítica de datos para dar respuesta a preguntas específicas. En la analítica de datos se usan herramientas estadísticas, desde las cuales se aplican metodologías capaces de interpretar conjuntos de datos con características particulares. Un instrumento útil para tales procesos es el *aprendizaje automático* —el cual se explicará más adelante—, que descubre o aprende patrones de manera autónoma a partir de un conjunto de datos. Mediante el aprendizaje automático es posible que los sistemas aprendan reglas, identifiquen patrones y tomen decisiones sin la necesidad de programar esas tareas de forma explícita —tarea que puede ser muy desgastante y, en algunos casos, impracticable—.

La operatividad del ciclo de vida de análisis de datos involucra diferentes perfiles técnicos y profesionales, entre los que se destacan los ingenieros de datos, los analistas de datos y los científicos de datos. Estos roles suelen trabajar juntos y, acompañados de profesionales con otros perfiles y experiencias, formar parte de equipos interdisciplinarios preparados para comprender problemáticas de carácter público y privado, planificar los proyectos según los datos disponibles y necesarios para desarrollarlo; explorar y descubrir patrones en los datos; elaborar los modelos descriptivos, predictivos y prescriptivos, y analizar los resultados obtenidos para determinar su utilidad y la mejor forma de aprovecharlos.

Dentro de las principales actividades que cumplen los equipos de ciencia de datos se encuentran las siguientes:

1. Buscar y consolidar distintas fuentes de información que puedan ser útiles para el desarrollo de un proyecto.
2. Depurar, cruzar y adecuar grandes cantidades de datos de diversos tipos.
3. Explorar y visualizar datos para descubrir valor y obtener percepciones e intuiciones a partir de los datos.
4. Utilizar lenguajes de programación y librerías especializadas para aplicar técnicas estadísticas, matemáticas y de aprendizaje automático, con el objetivo de obtener el mayor valor a partir de los conjuntos de datos disponibles.
5. Desplegar las herramientas desarrolladas y presentar los hallazgos a partir de los análisis a tomadores de decisiones y otros miembros de las entidades, de manera que puedan aprovecharse los resultados obtenidos.

1.1.1.3.1. *Técnicas y tipos de análisis de datos*

Como ya se mencionó, el análisis de datos es el proceso mediante el cual se transforman los datos en información comprensible que permita tomar decisiones basadas en datos. Los tipos de análisis de datos son diversos y su aplicación depende de las problemáticas y respuestas que se desee abordar (figura 1-4). Específicamente, hay 4 tipos de análisis de datos, a saber: descriptivo, de causalidad, predictivo y prescriptivo. La elaboración de esos cuatro tipos de análisis depende además de la clase de respuesta que se aspira adquirir del problema identificado previamente, de las habilidades y conocimientos con los que cuente el equipo de ciencia de datos, así como del nivel de madurez de analítica de una entidad o empresa.

Análisis descriptivo - *¿Qué es lo que sucede? - ¿Qué es lo que sucedió?*

Incorpora técnicas estadísticas para comprender el contexto actual de determinado problema, utilizando datos con distintos niveles de desagregación. También se emplean técnicas de visualización para generar gráficas y otras ayudas visuales que ayuden a comprender mejor la información.

Análisis causalidad - *¿Por qué sucede?*

Incorpora el uso de técnicas de estadística y econométricas de inferencia causal para efectuar análisis de causa y efecto, y comprender la forma como un conjunto de variables afecta una problemática.

Análisis predictivo - *¿Qué pasará? ¿Qué puede suceder?*

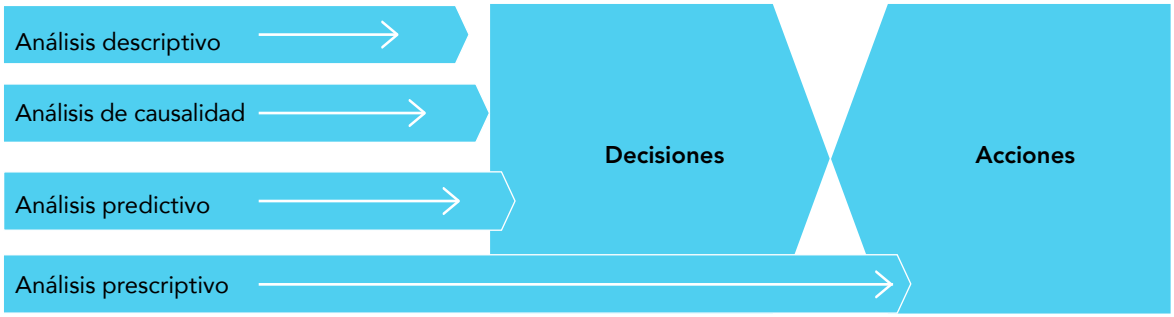
Incorpora modelos estadísticos y técnicas de aprendizaje de máquina. Este análisis permite predecir, por medio de la estimación de un valor o una probabilidad, las tendencias y comportamientos de un fenómeno, utilizando patrones históricos y condiciones actuales.

Análisis prescriptivo - *¿Qué se debería hacer?*

Incorpora algoritmos de optimización, análisis de decisión multicriterio y reglas de negocio para determinar cuál es la mejor acción (actual o futura) que debe tomar un agente en un caso o escenario determinado. Los modelos y las herramientas prescriptivas pueden utilizar como insumo los resultados de los otros tres tipos de análisis ya mencionados. En los análisis prescriptivos es importante la intervención del criterio humano, teniendo en cuenta la necesidad de incorporar un análisis ético a cualquier tipo de medida que se procure realizar.

En la figura I.1-4 se muestran los cuatro tipos de análisis de datos que se utilizan para abordar una problemática. Como se presenta en la figura I. 1-4, únicamente el análisis prescriptivo tiene la capacidad de plantear recomendaciones sobre acciones para abordar el problema o fenómeno.

Figura I.1-4. Tipos de análisis de datos



Fuente: elaboración propia adaptada de la Comisión Europea, 2016.

1.1.2. ANÁLISIS DE *BIG DATA* PARA LA TOMA DE DECISIONES BASADAS EN EVIDENCIA EN EL SECTOR PÚBLICO

De acuerdo con McKinsey Global Institute (2010), el desarrollo e implementación de sistemas y análisis de *big data* tiene efectos positivos tanto para las entidades públicas como privadas. El estudio establece que mediante la aplicación de la analítica de datos las entidades pueden mejorar y aumentar los niveles de generación de valor; por ejemplo, el sector de la salud en Estados Unidos podría capturar al año 2010, más de 300.000 millones de dólares al año, derivados del mejoramiento en los sistemas de eficiencia como producto de la aplicación de procesos de analítica de datos. Con base en la proliferación de los datos en la última década la cifra actual sería considerablemente mayor.

Los beneficios que se pueden alcanzar a partir del *big data* son diferentes para el sector público con respecto al sector privado teniendo en cuenta su naturaleza y misionalidad. En el caso de las entidades privadas los fines están enfocados principalmente en alcanzar mayores retornos económicos, mientras que en el sector público el proceso de creación de valor está orientado al cambio social y a la generación de valor público (Moore, 1995).

Para este contexto, el valor público se define como el valor creado por el Estado a través de las acciones que dan respuesta a las necesidades o demandas sociales (DAFP, 2016). Este objetivo se logra a través de la prestación de bienes y servicios, la participación ciudadana, el impulso y desarrollo de ciudades inteligentes[4] y la gestión del ciclo de políticas públicas —desde el diseño, implementación y evaluación— que responden a problemáticas de interés general. En el marco de creación de valor público, el análisis de datos se incorpora como un insumo que permite apoyar los procesos tanto operativos como misionales de las entidades para satisfacer las necesidades sociales.

4. *Las ciudades inteligentes* giran en torno a iniciativas que utilizan la innovación digital para prestar servicios de modos más eficientes y, por lo tanto, aumentar la competitividad general de una comunidad. (OCDE, 2020)

Frente a la prestación de servicios a la ciudadanía, los Gobiernos están incorporando la analítica de datos para mejorar los servicios existentes y aprovechar los conjuntos de datos novedosos para impulsar un servicio público completamente nuevo, ajustado a las necesidades particulares de ciudadanos y grupos de población específicos. Tal es el caso de la municipalidad de Córdoba, ciudad argentina que en el marco de la iniciativa Manos en la Data patrocinada por CAF, utilizó conjuntos de datos georreferenciados con el fin de hacer más eficiente el sistema de transporte público de Córdoba (CAF, 2019); por ello, diseñó una herramienta de movilidad a partir de técnicas estadísticas para detectar, en tiempo real, las brechas entre el recorrido programado de los buses micro y el tiempo efectivo del recorrido, y detectar incumplimientos en tiempos de frecuencias y servicios prestados.

La analítica de datos también aumenta la participación ciudadana, pues a través del aprendizaje automático de las redes sociales los Gobiernos pueden ser más receptivos al sentimiento y comportamiento ciudadano; por ejemplo, después de implementar alguna acción gubernamental. Además, mediante la gestión y analítica de datos es posible aumentar la disponibilidad de datos públicos estratégicos y de calidad para que los ciudadanos tomen decisiones.

Los procesos de explotación y uso de datos tienen como fin último mejorar el bienestar de los ciudadanos. Conviene anotar que el rol central de los ciudadanos no solo es el de usuarios sino también el de proveedores de los datos que captura y transforma el Gobierno después como información para la toma de decisiones. Por esa misma razón, resulta relevante la transparencia del sector público con respecto al tratamiento que da a los datos de las personas y a la creación de mecanismos efectivos de participación para que la ciudadanía pueda hacer uso de los datos y de las soluciones gubernamentales.

Frente al desarrollo de territorios y ciudades inteligentes, el análisis de *big data* y datos abiertos ofrece un gran potencial para la prestación de servicios urbanos, en donde la recopilación y tratamiento de datos de los ciudadanos, especialmente a través de Internet de las cosas (IdC), es fundamental para ofrecer mejores servicios y mejorar la calidad de vida de los ciudadanos.
Algunas de las aplicaciones del *big data* en ciudades inteligentes es ayudar a identificar zonas de la ciudad que requieren una mejor prestación de servicios de basura, agua y alcantarillado, identificar los patrones de congestión de tráfico y accidentes de tránsito, identificar patrones para proveer de manera eficiente la prestación de energía eléctrica, entre otros (Fundación Telefónica, 2016).

Por último, para acompañar el ciclo de las políticas públicas en cada una de sus etapas, los hacedores de políticas en todo el mundo están apoyando su diseño e implementación como fuente de creación de valor público en el análisis de imágenes satelitales, datos de movilidad y otros, para producir estudios detallados e indicadores tanto económicos como sociales alternativos y complementarios a las encuestas y los registros administrativos.

Un ejemplo de la aplicación de *big data* descrita es el uso de la analítica de grandes cantidades de datos para apoyar la lucha de la seguridad alimentaria en el mundo. Esta apuesta fue incorporada por la Organización Mundial de la Salud (OMS) en la plataforma de inocuidad de alimentos FOSCOLLAB dentro de una estrategia para incorporar fuentes de datos

de distintas disciplinas para soportar el análisis y la mitigación de riesgos frente con la seguridad alimentaria en el mundo.

Esta plataforma se nutre de datos de Internet y redes sociales, fuentes estructuradas como la base del Sistema Mundial de Monitoreo del Medio Ambiente o bases de datos oficiales de registros meteorológicos.

El almacenamiento de los datos estructurados y no estructurados y su posterior procesamiento utilizando aprendizaje automático y sistemas de recomendaciones permiten mejorar la planeación y diseño de estrategias para abordar diferentes dimensiones relacionadas con la seguridad alimentaria en el mundo (Yanseen, Bouzembrak, Hendriksen, & Staats, 2017).

1.1.2.1. El ciclo de las políticas públicas basadas en la evidencia y su relación con el *big data*

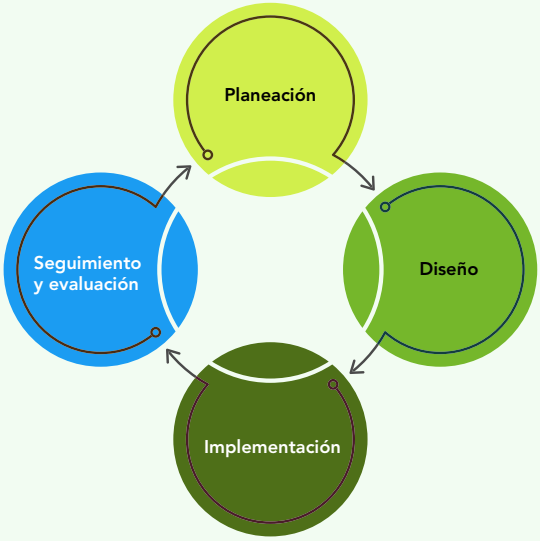
La formulación y evaluación de políticas públicas basadas en la evidencia es un enfoque que se viene adoptando en los últimos años y ha suscitado debate en las áreas de la economía y las ciencias sociales. Busca reestructurar los procesos de política pública al ubicar como idea principal para su ejecución la toma de decisiones de manera probatoria, reduciendo al máximo los análisis de carácter intuitivo o experimentales, a fin de disminuir el riesgo en la identificación del problema público, en la focalización de la intervención y en el desconocimiento del entorno en el que se formula la política (Studinka & Guenduez, 2019).

Las políticas basadas en la evidencia tienen como eje central los datos en la medida en que se requiere el uso y análisis regular de estos en el diseño, la implementación y la evaluación de políticas. Incluso se ha propuesto el concepto de *análisis de políticas* como un nuevo marco para abordar el nuevo paradigma de las políticas públicas (Marchi, Lucertini, & Tsoukias, 2016).

Por lo anterior, de acuerdo con el Banco Interamericano de Desarrollo, es fundamental que los países interesados en implementar este enfoque cuenten con procesos eficientes y efectivos para la recopilación, el almacenamiento, los análisis y el procesamiento de datos para el uso regular de datos. Para ello, se requiere el compromiso de los Gobiernos para evaluar de manera crítica y documentada sus intervenciones y la elaboración de una política de datos que habilite los requerimientos constantes de datos y de información (Banco Interamericano de Desarrollo, 2019).

Debido a lo anterior, en el presente documento el análisis del *big data* en las políticas públicas se orienta al enfoque de las políticas basadas en evidencia enmarcadas en el proceso o ciclo de política compuesto por las fases planeación, diseño, implementación, seguimiento y evaluación (figura I.1-5). Para cada una de las etapas, el análisis y la explotación de datos se orienta a un objetivo diferente con base en el propósito específico propio de cada etapa. A continuación, se describe brevemente la utilidad de la analítica considerando el ciclo de políticas públicas.

Figura I.1-5. Ciclo de las políticas públicas basadas en evidencia



Fuente: elaboración propia.

Fase de Planeación

Esta fase incorpora entre sus actividades la definición de una agenda de trabajo, la identificación del problema, las discusiones de política alrededor del problema y la participación de diversos actores.
El análisis de datos puede apoyar la identificación de problemáticas que la ciudadanía esté priorizando dadas sus necesidades; por ejemplo, con un análisis de redes sociales o algoritmos para análisis de texto que agrupan opiniones por temáticas y problemáticas. También puede ayudar a identificar dinámicas sociales entre actores inmersos en la elaboración de la política que pueden ejercer un rol promotor o renuente frente una intervención determinada.

Fase de Diseño

Incluye actividades como definir de objetivos, alternativas de intervención, acciones y metas para el posterior seguimiento. Los tomadores de decisiones de política requieren de información para diagnosticar la problemática que se aspiran intervenir, tener un panorama completo del estado del arte.

En este punto es pertinente efectuar un análisis descriptivo para conocer la situación actual, y predictivo para estimar los efectos de la intervención de política pública. La disponibilidad de datos de registros administrativos, encuestas y otras fuentes de información estructuradas y no estructuradas fortalece la evidencia del problema para aproximarse con más precisión en el diseño de la política.

Son ejemplos de cómo la analítica de datos puede aportar a esta fase los siguientes:

1. La utilización de datos satelitales para detectar déficits de infraestructura urbana —como la detección del tamaño, forma y expansión de asentamientos informales—;

2. La utilización de microdatos de empleados y empleadores para caracterizar los patrones en la movilidad laboral y construir herramientas de formulación de políticas de empleo (Studinka & Guenduez, 2019).

Fase de Implementación

En esta etapa se ejecutan las acciones para alcanzar los objetivos de la política definidos en la etapa de diseño. La implementación de la política crea por sí misma los datos que luego se usarán para la evaluación y el seguimiento. La generación de datos en tiempo real puede apoyar a los Gobiernos en evaluaciones de corto plazo que les permitan identificar si las acciones están teniendo los efectos esperados. En este sentido, conforme a los datos disponibles, será posible ajustar el diseño y la implementación de la política pública.

Son ejemplos de cómo la analítica de datos puede aportar a esta fase los siguientes:

1. Detectar fraude en la entrega de recursos o apoyos en especie para cierta población beneficiaria de un programa;

2. Modificar zonas de cobertura de policía por cuadrante en un programa de seguridad ciudadana, a partir de la identificación de dinámicas de inseguridad ciudadana, que se han intensificado por una situación de coyuntura dada;

3. Usar grandes datos para monitorear efectos adversos de medicamentos entregados a una población (Studinka & Guenduez, 2019).

Fase de Evaluación

Durante esta etapa se analiza el impacto previsto de la política sobre los resultados esperados. La evidencia permite estimar los beneficios alcanzados por la política, e identificar las posibles modificaciones que requiera hacérsele, en relación con los resultados de la política. Por lo general, la fase de evaluación de la política se lleva a cabo una vez esta ha finalizado; sin embargo, el principal aporte del *big data* en esta etapa es que permite la disponibilidad y el análisis de datos en tiempo real, para adelantar evaluaciones de impacto con mayor rapidez que las evaluaciones de impacto tradicionales. Lo anterior significa que no es necesario esperar a la ejecución completa de la política para reconocer fallas o éxitos en su implementación (Schöllhammer, Parycek, & Höchtl, 2016).

La contribución más destacada del procesamiento de datos en tiempo real es que los resultados de la evaluación de políticas están disponibles en el mismo momento en que se generan los datos. Así ocurre la transformación del ciclo de políticas públicas orientada hacia la evaluación continua (Schöllhammer, Parycek, & Höchtl, 2016).

Algunos ejemplos de cómo la analítica de datos puede aportar a esta fase son los siguientes:

1. La evaluación del efecto de una nueva política tributaria en tiempo real para determinar si tiene o ha tenido el efecto deseado;

2. La aplicación de metodologías de *machine learning* para estimar buenos contrafactuales que permitan mejorar el análisis entre grupo de control y tratamiento;

3. El incremento de fuentes de información que complementen las estadísticas censales y de encuestas nacionales que a menudo pueden estar desactualizadas.

La aplicación de analítica de datos para apoyar el ciclo de las políticas públicas exige la definición de una infraestructura de datos y de explotación y de una analítica de datos que habiliten las condiciones para que los países fortalezcan la toma de decisiones basadas en la evidencia. Como se expondrá en el capítulo siguiente, el marco normativo de política es fundamental para eliminar las barreras que reducen la disponibilidad de los datos y la falta de disponibilidad de talento humano para su aprovechamiento.

Adicional al tema de datos, es necesario que los países incorporen una gobernanza de las políticas basadas en la evidencia, que promueva la destinación de presupuesto y el fortalecimiento de las capacidades nacionales para el aprovechamiento de datos. En la próxima sección se mostrará la apuesta del Gobierno de Colombia durante los últimos años para definir un marco de política que habilite el aprovechamiento de datos para la generación de valor público y social.

1.1.3. CASOS DE ÉXITO EN EL USO DE ANALÍTICA DE DATOS Y *BIG DATA* PARA ABORDAR PROBLEMÁTICAS DE CARÁCTER PÚBLICO

El uso de analítica de datos para abordar problemáticas de carácter público es cada vez más frecuente entre los distintos gobiernos en todo el mundo. A su vez, los organismos multilaterales como el Banco Interamericano de Desarrollo (BID), CAF, la Organización Mundial de las Nacionales Unidas (ONU), entre otras, promueven, desde distintas estrategias en alianza con los sectores público y privado, la gestión, el uso y el aprovechamiento de datos para la toma de decisiones en el marco de los problemas de carácter público. En general, estas iniciativas destacan la utilidad que tiene la ciencia de datos para ayudar a mejorar vidas, utilizando datos abiertos, datos del sector privado a través de mecanismos de cooperación e intercambio efectivo y eficiente de datos con las entidades públicas.

La OCDE ha identificado cuatro áreas de desarrollo que se benefician en mayor medida con el uso de *big data* (Banco Mundial, 2014):

1. La estimación y el análisis sociodemográficos,

2. El crecimiento económico, la innovación y la investigación,

3. El análisis social, la vulnerabilidad y la resiliencia ambiental,

4. El análisis y la detección de riesgos de salud pública.

Sobre el último, recientemente la emergencia mundial ocasionada por la COVID-19 dejó en evidencia la utilidad de los datos para abordar problemáticas complejas derivadas de la pandemia, no solo en términos de diagnóstico y pronóstico de las tendencias de contagio, sino también como activo para la implementación de estrategias de reactivación social y económica.

Para analizar información en tiempo real sobre lo que sucede con las personas en términos de salud, desplazamientos, y necesidades de consumo, los datos de encuestas y registros administrativos se han complementado con información de telefonía móvil, redes sociales, búsquedas en Google. Gran parte de los datos los generan las empresas privadas, lo que ha puesto en evidencia la necesidad de fortalecer el intercambio de datos entre el sector público y el sector privado, sin dejar de proteger los derechos de los ciudadanos en materia de privacidad y seguridad de la información, así como los intereses privados de las empresas.

El trabajo colaborativo entre el sector público y el sector privado, lo mismo que la compartición de los datos conforme al marco normativo aplicable a su protección, ha traído importantes resultados a en el ámbito internacional, para identificar factores de riesgo de contagio, pronosticar las dinámicas de contagio por zonas geográficas y tipos de población, tomar decisiones para la reapertura económica, identificar interrupciones en las cadenas de suministros, predecir el comportamiento del mercado laboral y para otras temáticas. Por ejemplo, la iniciativa *National Data Sharing Partnership to Fight COVID-19* es un esfuerzo por recopilar, intercambiar y analizar de manera colaborativa, grandes conjuntos de datos que contribuyan a la comprensión de la enfermedad, los efectos del tratamiento, la identificación de fármacos, entre otras (Center for Leading Innovation & Collaboration, 2020).

El Foro Económico Mundial utilizó técnicas de modelamiento y simulación basados en agentes para analizar el comportamiento de las personas ante las medidas de confinamiento. Los datos móviles fueron fundamentales para identificar los patrones de movilidad y el número de kilómetros que recorrieron las personas en Estados Unidos. Lo anterior permitió medir la efectividad de las medidas de los Gobiernos para que sus ciudadanos se mantuvieran en casa. Otro de los modelos diseñados por el Foro Económico Mundial fue un modelo integrado de simulación de detección de demanda, interrupción de la cadena de suministro y planificación de la fuerza laboral, el cual incorporó varios silos de toma de decisiones en un solo modelo dinámico, en el que también se integraron nuevas variables para modelar escenarios de recuperación macroeconómica de la pandemia (Foro Económico Mundial, 2020).

Por otra parte, el Banco Interamericano de Desarrollo creó un portal interactivo para revelar la movilidad de 22 países de América Latina y el Caribe, para medir el nivel de eficiencia de las medidas de confinamiento dadas por los Gobiernos de la región. El aplicativo se alimenta de datos georreferenciados de teléfonos celulares; también se han utilizado análisis de redes sociales para identificar la reacción de los ciudadanos ante a las acciones implementadas por los gobiernos (BID, 2020). Los anteriores ejemplos demuestran el potencial de recopilar datos en tiempo real para analizar diversas situaciones de carácter público y para los hacedores de políticas públicas en todo el mundo, quienes reconocen la necesidad de gestionar los riesgos derivados del uso de los datos.

Específicamente en Colombia, se creó el Programa de Ingreso Solidario para atender a la población vulnerable que se afecta mayormente por las consecuencias de la pandemia de la COVID-19, a partir de un trabajo colaborativo entre el sector público y el sector privado. Los datos de telefonía móvil, los registros administrativos de seguridad y programas sociales, junto con los datos del sector financiero permitieron diseñar e implementar ese programa de transferencias económicas del Estado en tiempo récord.

Por otra parte, en el marco de la iniciativa Manos en la Data-Colombia, coordinada de manera conjunta entre CAF, Alianza CAOBA y DNP, se elaboraron seis soluciones de analítica de datos, que permitieron la comprensión y análisis de problemáticas de salud, seguridad ciudadana, transporte de carga carretero, entre otros, dentro del marco de las situaciones derivadas de la emergencia sanitaria de la COVID-19. Los resultados de la iniciativa se presentan en el tercer capítulo de este documento.

Sin bien la importancia de los datos se ha visibilizado con mayor fuerza durante la coyuntura social y económica en estos tiempos de pandemia, la analítica de datos ha acompañado la gestión de soluciones de interés público desde hace varios años. En lo relacionado con los Objetivos de Desarrollo Sostenible (ODS)[5], el *big data* se convierte en una oportunidad para apoyar la implementación de acciones que deriven en su cumplimiento y contribuyan a la fase de seguimiento y evaluación de los logros. Como ya se expuso, la consecución de los resultados de proyectos y políticas demandan la existencia de métodos de recolección, generación y análisis de datos para medir y evaluar información fundamental en la toma de decisiones y, por otra parte, el análisis de datos mejora la prestación de servicios y productos adaptados a las necesidades reales de la población. Por los tanto, el análisis de datos contribuye positivamente al logro de los 17 ODS a escala global. Los proyectos que se describen de manera muy general en la tabla I.1-2 son ejemplos de cómo el aprovechamiento de los datos ha abordado diferentes problemáticas de interés público.

5. Surgen como respuesta a los problemas primordiales de las naciones y buscan mejorar la calidad de vida de sus habitantes.

Tale I.1-2. Recopilación de proyectos en el marco de los Objetivos de Desarrollo Sostenibles

LOGO DEL ODS	NOMBRE DE PROYECTO Y RESEÑA
	**Cartografía de la pobreza por satélite (Sri Lanka)** Se utilizan indicadores de datos satelitales de alta resolución para obtener información sobre factores regionales, tales como el nivel de urbanización, la infraestructura y los recursos naturales. La iniciativa ha ayudado a elaborar mapas de pobreza locales más precisos que facilitan la orientación de políticas de desarrollo (World Bank Group, 2016).
	**Estimación de los precios de los alimentos a través de las señales de las redes sociales (Indonesia)** Se clasifican datos generados por medio de redes sociales, utilizando palabras clave y frases relacionadas con algunos productos alimentarios específicos, con el fin de obtener información sobre el comportamiento diario de sus precios. El modelo elaborado permite hacer seguimiento del precio de los alimentos en tiempo real, lo cual es útil para la formulación de políticas y la gestión de riesgos económicos (UN Global Pulse, 2014).
	**"Nos sentimos bien" (Brasil)** Es una herramienta algorítmica de análisis predictivo que permite hacer seguimiento y monitoreo de la frecuencia cardíaca, el pulso y la respiración de los bebes recién nacidos. Es posible predecir de manera temprana las infecciones que adquieren los neonatos antes de que presenten algún síntoma por el que requieran recibir atención oportuna (Health Ecosphere, 2020).

LOGO DEL ODS	NOMBRE DE PROYECTO Y RESEÑA
	**Predicción de la deserción escolar utilizando datos administrativos (Guatemala y Honduras)** El proyecto utiliza datos administrativos del sector educativo a fin de reducir las altas tasas de deserción escolar. Se elaboran modelos de predicción para identificar a los estudiantes que presentan la mayor probabilidad de abandonar sus estudios antes de culminarlos, lo cual permite desarrollar un sistema de alerta temprana (World Bank Group, 2017).
	**Análisis de actitudes hacia la anticoncepción y el embarazo adolescente utilizando datos de redes sociales (Reino Unido)** Se usaron datos digitales en tiempo real para analizar la percepción de los ugandeses sobre el embarazo en la adolescencia y los métodos anticonceptivos. Se construyó un tablero que permitió hacer seguimiento mensual a la percepción de los ciudadanos frente a lo relacionado con la planificación familiar (UN Global Pulse, 2014).
	**Extrayendo información de las redes sociales para entender las percepciones públicas (Global)** Se utiliza una taxonomía de palabras claves buscando identificar las percepciones generales de las personas sobre los sistemas de saneamiento, tal como la expresan en las redes sociales. La iniciativa provee indicadores para monitorear y evaluar los cambios en las percepciones de la gente durante las campañas de salud pública (UN Global Pulse, 2014).
	**Medidor inteligente de datos de electricidad (Reino Unido)** En este proyecto se diseñó un algoritmo para mapear las medidas de consumo como variables proxis de ingresos, utilizando estimaciones no paramétricas y aprendizaje automático estadístico. La herramienta permite identificar las discrepancias entre los ingresos imponibles representados y declarados para estimar el efecto causal del programa en los pagos de impuestos (*Big Data* UN Global Working Group, 2019).
	**Datos de portales de empleo en línea para examinar las políticas del mercado laboral (India)** Se aprovechan los datos que provee un portal de empleo en línea, para obtener información sobre las aptitudes que demandan los empleadores en el mercado laboral y sobre las características de los solicitantes de empleo. Este tipo de iniciativas aportan en la formulación y el perfeccionamiento de las políticas del mercado laboral y contribuyen a cerrar la brecha entre los empleadores y los buscadores de empleo (World Bank Group, 2017).
	**Plataforma "*Open Trafic*" (Filipinas)** Se aprovechan el crecimiento del uso de teléfonos inteligentes y la aparición de empresas de aplicaciones de transporte urbano para obtener datos sobre el tráfico en tiempo real. Se crea la plataforma de código abierto "Open trafic" que permite recoger, visualizar y analizar este tipo de datos, con ello se reducen los costos de recolección, codificación y análisis respecto a los métodos tradicionales (World Bank, 2015).
	***Big data* para la inclusión Financiera (Ghana)** Se establece un perfil estadístico de los usuarios de servicios financieros digitales por medio de datos e investigaciones socioeconómicas en África. Posteriormente, se identificaron perfiles coincidentes entre los clientes telefónicos que no utilizaban esos servicios, como potenciales usuarios de servicios financieros digitales. Este tipo de proyectos ayudan a extender la cobertura de los servicios bancarios (World Bank Group).

LOGO DEL ODS	NOMBRE DE PROYECTO Y RESEÑA
	**Proyecto: EXTREMA (Unión Europea)** El Servicio de Alertas de Temperatura EXTREMA proporciona información a los individuos, por medio de una aplicación móvil, sobre su riesgo personalizado a las temperaturas extremas a las que puede llegar a estar expuestos según su localización. La iniciativa promueve la adopción de conductas de autoprotección; los Gobiernos locales se han apoyado en ella para para reforzar sus acciones de preparación y respuesta ante emergencias climáticas (ICLEI, 2019).
	**Utilizando "*big data*" para apoyar la gestión de los desechos electrónicos (China)** Se diseña una aplicación web que busca recolectar y reciclar los dispositivos electrónicos que los usuarios quieren desechar. La aplicación permite mejorar la supervisión y el comportamiento de reciclaje en la eliminación de los desechos electrónicos, y logra concientizar sobre los enfoques ambientalmente apropiados para ente tipo de procesos (UNDP China, 2015).
	**Sistema integrado de medio ambiente (Singapur)** Se diseñó una plataforma que permite la recopilación centralizada en tiempo real de datos del medio ambiente y fenómenos meteorológicos. A su vez, posibilita mejorar las capacidades de predicción para identificar alertas tempranas ante catástrofes ambientales (*Big Data* UN lobal Working Group, 2019).
	**Monitoreo de santuarios de tiburones con datos satelitales (Islas Marshall)** Con ayuda del Observatorio Pesca Global se ha podido monitorear la actividad de las embarcaciones dentro de los límites de los santuarios de tiburones en las en las Islas Marshall. Las plataformas del observatorio proveen datos recolectados por medio de un sistema de identificación automática por satélite que permite visualizar, rastrear y aportar transparencia a la actividad pesquera (Society for Conservation Biology, 2019).
	**Herramienta para mejorar el acceso a datos geoespaciales sobre los ecosistemas (Zimbabue)** Se recopilaron diferentes datos georreferenciados por categorías relevantes para la biodiversidad, con el objetivo de desarrollar una herramienta que permita el acceso amigable a esa información. La herramienta ha servido para apoyar la formulación de las estrategias y planes de acción nacionales para la conservación, preservación y protección de la biodiversidad (UN Global Pulse, 2015).
	**Artemisa (Canadá)** Este proyecto de investigación analiza el sentimiento ciudadano sobre las instituciones estatales, la confianza en el gobierno y su relación con los disturbios ciudadanos. El análisis de sentimientos se lleva a cabo a partir de datos de las redes sociales durante un período de un año (Banco Mundial, 2015).
	**"*Qatalog*" (global)** *Qatalog* es una herramienta que permite extraer información útil de grandes bases de datos provenientes de diversas fuentes, como la radio o las redes sociales, en cualquiera de los 34 idiomas disponibles. La plataforma busca diseñar y escalar sistemas de inteligencia colaborativa con la retroalimentación de los usuarios. Este tipo de iniciativas movilizan los programas de investigación de la comunidad académica internacional (UN Global Pulse, 2018).

Fuente: elaboración propia.

Como se evidencia en la tabla I. 1-2, el uso de datos masivos y analítica de datos se ha utilizado para abordar diferentes problemáticas de interés público y social para mejorar la calidad de vida de las personas.

De acuerdo con el Grupo Asesor de Expertos Independientes sobre la Revolución de Datos para el Desarrollo Sostenible (GAEI), para aprovechar al máximo los datos para el desarrollo social y económico, es necesario que los países adopten principios básicos (GAEI, 2014), a saber:

1 

La calidad e integridad de los datos

Es decir que todo el proceso de diseño, recopilación, análisis y difusión de los datos deber de alta calidad.
Lo anterior para evitar incurrir en conclusiones o análisis erróneos por causa de datos de mala calidad.

2 

La pertinencia del ciclo de vida del análisis de los datos con el ciclo de las políticas públicas

El valor de los datos se incrementa en la medida en que sean datos actualizados que estén disponibles para adoptar decisiones.

3 

La transparencia y la apertura de los datos

Considerando los elementos técnicos y jurídicos para su consumo.

4 

La protección y privacidad de los datos

Es fundamental para garantizar la confianza de la ciudanía en relación con los datos que se recopilan y se tratan.

5 

La gobernanza de los datos

Para garantizar una adecuada gestión de su ciclo de vida, especialmente en aspectos como la calidad y la estandarización.

6 

El fortalecimiento de las capacidades de los organismos

Públicos en términos de infraestructura tecnológica, y capital humano para aprovechar los datos en sus diversas fuentes de información.

Dado el potencial que tienen los datos, es importante reconocer los esfuerzos que ha adelantado Colombia para impulsar su aprovechamiento en la toma de decisiones en cual ha tenido en cuenta los seis principios mencionados.

En el capítulo I. 2 se hace un recuento de las estrategias que ha implementado Colombia para aumentar el nivel de datos disponibles y mejorar su utilidad, alcanzar un mayor grado de apertura y transparencia, mejorar la capacidad de las personas para utilizar los datos y fortalecer el marco jurídico para su compartición.

CAPÍTULO 1.2

APUESTAS DE POLÍTICA PÚBLICA **PARA IMPULSAR EL APROVECHAMIENTO DE DATOS** EN COLOMBIA

LA *POLÍTICA DE GOBIERNO EN LÍNEA* (PGL) ADELANTADA POR EL GOBIERNO DE COLOMBIA DESDE EL AÑO 2008 (DECRETO 1551 DE 2008) FUE EL PUNTO DE PARTIDA PARA SENTAR LAS BASES DE LA TRANSFORMACIÓN DIGITAL DEL ESTADO COLOMBIANO. LA ESTRATEGIA BUSCÓ A TRAVÉS DE LA APROPIACIÓN DE LAS TIC GENERAR INCLUSIÓN SOCIAL Y AUMENTAR LA COMPETITIVIDAD DEL PAÍS Y MEJORAR LA EFICIENCIA Y TRANSPARENCIA DEL ESTADO A TRAVÉS DE LA CONSTRUCCIÓN DE UN GOBIERNO ELECTRÓNICO.

En ese contexto, conviene señalar que aunque el aprovechamiento de los datos no era un componente directo de la PGL, como se mostrará más adelante, durante los siguientes cinco años se consolidaron las bases para estructurar la política de explotación de datos en el país, que se diseñaría e implementaría en una etapa posterior. A partir de la Política de Gobierno en Línea, las TIC se consolidaron como un elemento fundamental para aumentar la eficiencia y la eficacia de la gestión pública, aumentar la participación ciudadana y promover la transparencia del sector público.

Durante esa fase se materializaron esfuerzos para estructurar un marco normativo orientado a la protección de datos personales, el acceso a la información pública y el intercambio de información entre entidades del Estado. En el año 2010, se publicó el Decreto 235 que estableció las pautas de intercambio de datos entre entidades que permitan el cumplimiento de sus funciones públicas, a través de mecanismos electrónicos o telemáticos para compartir información en el marco de observancia del derecho fundamental a la intimidad.

En materia de regulación de protección de datos personales, se aprobó la Ley 1581 de 2012, en la que se incluyeron las garantías para los titulares de los datos en materia de acceso, uso, actualización y rectificación de los datos personales y se definió la clasificación de la información de los datos privados de carácter público, semiprivado, privado y sensible. Además, la Ley 1712 de 2014 reguló el derecho al acceso a la información pública y definió los principios aplicables para su implementación, así como las excepciones a la publicación de información. A partir de esta ley se impulsó en mayor medida el principio de transparencia del Gobierno a partir de la publicación de

datos abiertos, para su reutilización y uso por parte de los ciudadanos, otras entidades públicas, el sector privado y la academia.

Durante la primera fase de la política, el objetivo principal consistió en la consolidación de un Estado más transparente, participativo y eficiente; en la segunda fase, la estrategia se enfocó en la prestación de servicios de colaboración con la sociedad; por último, en 2014, la política se concentró en un Estado Abierto, transparente y participativo a partir del uso de las TIC. En esta fase la estrategia se enfocó en los siguientes cuatro componentes: TIC para servicios, TIC para el Gobierno abierto, TIC para la gestión y seguridad y privacidad de la información. Las estrategias implementadas por Gobierno en Línea se resumen en la tabla I.2-1.

Tabla I.2-1. Fases de la política de Gobierno en Línea

ETAPA	MARCO LEGAL	PRINCIPIOS
De 2008 a 2012	Decreto 1151 de 2008	• Visión unificada del Estado • Acceso igualitario y multicanal • Protección de la información personal • Confianza y credibilidad del Gobierno
De 2012 a 2014	Decreto 2693 de 2012	• Construcción colectiva • Innovación • Neutralidad de red • Confianza y seguridad
De 2014 a 2018	Decreto 2573 de 2014	• Entrega destacada de servicios a los ciudadanos • Apertura y reutilización de los datos públicos • Estandarización • Interoperabilidad • Neutralidad de red • Innovación • Colaboración

Fuente: elaboración propia con base en OCDE (2014).

Como resultados de la Política de Gobierno en Línea se destacan los avances que tuvo el país en relación con la apertura y disponibilidad de datos abiertos. Por ejemplo, en 2017 se destaca el puntaje que alcanzó el país en el índice gubernamental de datos abiertos *(OurData Index* por sus siglas en inglés) que mide la disponibilidad de datos en el portal de datos nacional, el acceso a los datos públicos y el soporte activo para su reutilización. De acuerdo con los resultados, Colombia estuvo entre los cinco primeros países de la OCDE. De otra parte, Colombia ocupó la posición 24 entre 114 países en el Barómetro de Datos Abiertos de 2016.

Posteriormente en 2018, la estrategia de Gobierno en Línea se transformó en *Gobierno Digital* mediante el Decreto 1008. Este nuevo enfoque tiene como principal objetivo "Promover el uso y aprovechamiento de las tecnologías de la información y las comunicaciones para consolidar un Estado y ciudadanos competitivos, proactivos, e innovadores, que generen valor público en un entorno de confianza digital". Con base en las recomendaciones de la OCDE, la Política de Gobierno Digital definió su enfoque orientado a la generación de valor público; es decir, que con esta estrategia no solo se busca la entrega de productos y servicios TIC a

la ciudadanía sino también introducir el concepto de generación de valor cuando la ciudadanía utiliza esos productos y servicios (OCDE, 2014).

La Política de Gobierno Digital estableció tres habilitadores —arquitectura TI, seguridad y privacidad de la información, servicios ciudadanos digitales—, para alcanzar cinco propósitos:

1 Habilitar y mejorar la provisión de servicios ciudadanos digitales.

2 Lograr procesos seguros y eficientes a través de las capacidades de gestión de las tecnologías de la información.

3 Empoderar a los ciudadanos a través de un estado abierto.

4 Impulsar el desarrollo de territorios y ciudades inteligentes.

5 Tomar decisiones basadas en datos a partir del aumento del uso y aprovechamiento de la información.

En el marco de la Política de Gobierno Digital y siguiendo las recomendaciones de la OCDE para impulsar el aprovechamiento de datos en el sector público (OCDE, 2018), el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) ha establecido lineamientos y guías para que las entidades públicas implementen buenas prácticas en lo referente a la gobernanza de sus datos, la apertura de datos abiertos, la seguridad y privacidad de la información y la compartición de datos entre entidades.

Entre los principales aspectos identificados por la OCDE en su informe, se resaltó la necesidad de fortalecer y establecer una directriz de gobernanza de los datos para articular el marco normativo y de política con los estándares técnicos orientados a alcanzar objetivos nacionales relacionados con la cultura de datos en las entidades públicas, la interoperabilidad de datos, y el fortalecimiento de las capacidades de las entidades del Gobierno para gestionar de manera efectiva el ciclo de vida de los datos y promover su aprovechamiento.

1.2.1. CONSOLIDACIÓN DEL MARCO DE POLÍTICA PARA LA EXPLOTACIÓN DE DATOS Y *BIG DATA* Y LA TRANSFORMACIÓN DIGITAL DEL ESTADO

De manera paralela al avance de la Oolítica de Gobierno en Línea implementada por el país y el desarrollo exponencial del sector de tecnologías de la información y las comunicaciones, el Departamento Nacional de Planeación creó la Dirección de Desarrollo Digital (DDD), a partir del Decreto 2189 de 2017, teniendo en cuenta las demandas y necesidades del Gobierno para la promoción y desarrollo del sector TIC. Como una de sus funciones, la DDD está encargada del desarrollo y la ejecución de estrategias e iniciativas para la explotación e innovación basadas en datos.

La dinámica de generación de datos en todo el mundo y el valor de estos para alcanzar beneficios de carácter público y privado, hizo que en 2014 el Gobierno de Colombia, a través del Plan Nacional de Desarrollo 2014-2018, le otorgara al Departamento Nacional de Planeación (DNP) la responsabilidad de liderar la estrategia de *big data* en todo el país. Como desarrollo de dicho mandato, en 2018, la Dirección de Desarrollo Digital del DNP diseñó, en articulación con entidades del orden nacional, el Documento CONPES 3920: *Política Nacional de Explotación de Datos (big data)*. Así, Colombia se convirtió en el

octavo país del mundo y el primero en América Latina en tener una política para habilitar las condiciones para el análisis de *big data*.

En el Plan Nacional de Desarrollo 2018-2022: *Pacto por Colombia, pacto por la equidad*, se definió el Pacto VII: *Pacto por la transformación digital del país* mediante el cual "las tecnologías digitales se presentan como habilitadoras para la agregación de valor en la economía, generadoras de nuevos negocios y puerta de entrada a la industria 4.0" (DNP, 2019). La analítica de datos y el *big data* se mencionan como habilitadores transversales de diferentes pactos del en el PND2018-2022, y en lo relacionado con la infraestructura de datos públicos se la señala como uno de los 12 principios de la transformación digital en el país.

Con base en el contexto descrito, en 2019 se elaboró el Documento CONPES 3975: *Política Nacional para la Transformación Digital e Inteligencia Artificial* cuyo objetivo consiste en aumentar la generación de valor social y económico a partir de la transformación digital. La política establece la necesidad de continuar implementando las acciones del Documento CONPES 3920 para la consolidación de una infraestructura de datos públicos que habilite la transformación digital y la inteligencia artificial en el país.

1.2.2. RESULTADOS DEL DOCUMENTO CONPES 3920: *POLÍTICA NACIONAL DE EXPLOTACIÓN DE DATOS (BIG DATA)* PARA HABILITAR LAS CONDICIONES PARA EL APROVECHAMIENTO DE DATOS

La Política Nacional de Explotación de Datos (Documento CONPES 3920) está vigente hasta el año 2021; su objetivo principal es aumentar el aprovechamiento de datos en Colombia, a través de líneas específicas de acción que habilitan las condiciones del país, para aumentar la disponibilidad de datos públicos digitales, fortalecer el marco jurídico e institucional para la explotación de datos, aumentar el capital humano para la explotación de datos y acrecentar la cultura de datos en el país, no solo por parte de las entidades públicas sino también por parte de los ciudadanos y los actores privados.

De acuerdo con el diagnóstico efectuado para la elaboración de la política pública durante 2017, las entidades públicas del país enfrentaban importantes retos en materia de interoperabilidad, apertura de datos públicos, incertidumbre sobre la normativa vigente, y en general ausencia de condiciones para la implementación de proyectos explotación de datos. En lo tecnológico, únicamente el 26% de las entidades habían asignado presupuesto para llevar a cabo procesos de digitalización y el 67,3% de ellas tenía menos del 70% de información digitalizada. Otro de los problemas identificados fue la débil interoperabilidad entre las entidades del sector público; aunque el país disponía para ese entonces de un marco normativo para efectuarla desde el año 2005, el 54,5% de las entidades públicas manifestó tener procesos de interoperabilidad con menos de 3 entidades públicas.

Otros de los retos identificados era la falta de reconocimiento del valor de los datos como activo para la generación de valor social y económico por parte de las

entidades públicas. Según el Documento CONPES 3920, cerca de la mitad de las entidades públicas consideraba que aumentar la disponibilidad de los datos abiertos no solucionaría ninguna necesidad interna y el 31 % consideraba que no era relevante la compartición de datos con el sector privado. La ausencia de una cultura de datos en las entidades públicas dificulta la posibilidad de generar cambios estratégicos tanto organizacionales como tecnológicos para posicionar a los datos como un insumo fundamental para la producción de bienes y servicios que generen valor público.

El CONPES 3920 ha implementado acciones estratégicas para abordar los retos ya mencionados. En relación con el aumento de datos abiertos digitalizados, se definió e implementó un marco de interoperabilidad (*X Road*) para el intercambio de información entre entidades públicas, en el que se definieron los lineamientos políticos, técnicos y semánticos para hacer efectiva la compartición de datos.
Como complemento, se creó el portal de *Software* Público para disponer licencias de código abierto; además, se creó dentro del Portal de Datos Abiertos una herramienta de gestión de activos de información para las entidades públicas y se elaboraron guías para estandarizar los criterios de calidad de datos, registros administrativos y datos abiertos enlazados para las entidades públicas.

Con relación al fortalecimiento del marco institucional y normativo, se definieron no solo los principios sino también las necesidades para la implementar un marco de gobernanza de la infraestructura de datos públicos, así como las condiciones para su puesta en marcha cuyo requisito fundamental es la articulación de los roles y funciones de Presidencia de la República, el Ministerio de las Tecnologías de la Información y las Comunicaciones, la Superintendencia de Industria y Comercio, el Departamento Nacional de Planeación y otras entidades. También se definieron las exigencias jurídicas para la implementación de asociaciones público-privadas para la explotación de datos, la identificación de principios éticos aplicables a la explotación de datos y la definición de criterios que deben cumplir los proyectos de explotación de datos a fin de garantizar su transparencia.

Respecto a la línea de acción para aumentar el capital humano disponible para el aprovechamiento de datos, el Departamento Administrativo para la Función Pública introdujo en 2020 el eje de transformación digital en el Plan Nacional de Formación y Capacitación, en el cual se incluyeron temas de formación en explotación de datos para los funcionarios de las entidades públicas. Por otra parte, el MinTIC ha elaborado diferentes iniciativas para formar científicos de datos en el país, fortalecer las capacidades de los ciudadanos para el aprovechamiento de datos abiertos. Otra de las acciones para resaltar es el diseño de una plataforma Data Sandbox, como espacio colaborativo para que las entidades públicas puedan elaborar prototipos de analítica de datos y la elaboración del estudio de brechas de capital humano en el sector TIC con énfasis en las actividades de explotación de datos.

Entre las acciones llevadas a cabo para aumentar la cultura de datos en las entidades públicas, el Departamento Nacional de Planeación diseñó el *modelo de explotación de datos* como una herramienta de autodiagnóstico, con el fin de evaluar las capacidades de las entidades para la explotación de datos, e identificar de manera general el valor potencial que estas podrían generar a partir de la inversión en *big data*. Otra de las acciones para aumentar la cultura de datos fue la creación, en la Dirección de Desarrollo Digital, de la primera Unidad de Científicos de Datos del país con el propósito de estructurar un equipo de trabajo que apoyara a las direcciones técnicas del DNP y a otras entidades del Estado, para visibilizar el valor de la analítica y explotación de datos.

En el capítulo I.3 se describe la experiencia de la Unidad de Científicos de Datos del DNP y los resultados alcanzados en la elaboración de los proyectos de analítica de datos en el sector público; en sus secciones se exponen ejemplos de buenas prácticas y lecciones aprendidas, además de aportes a la transferencia del conocimiento y la apropiación de la explotación de datos en otras entidades públicas colombianas.

CAPÍTULO 1.3

EXPERIENCIA DE LA UNIDAD DE CIENTÍFICOS DE DATOS DEL DEPARTAMENTO NACIONAL DE PLANEACIÓN

1.3.1. ¿QUÉ HACE LA UNIDAD DE CIENTÍFICOS DATOS?

La Unidad de Científicos de Datos (UCD) del Departamento Nacional de Planeación es el primer equipo que se conformó en el Gobierno nacional para la explotación de datos en el sector público en Colombia. La UCD se encarga de la formulación, la ejecución y el seguimiento de proyectos con componente de analítica de datos, que permiten a las direcciones técnicas del DNP y otras entidades del Estado aumentar la creación de valor en sus procesos a través del aprovechamiento efectivo de sus datos, lo que se traduce comúnmente en una toma de decisiones objetiva. Para ello, el equipo cuenta con capacidades técnicas en análisis estadístico, aprendizaje de máquina (*machine learning*), análisis de texto y procesamiento de imágenes que le permiten trabajar con información no estructurada —es decir, textos, imágenes, audio y video—, y con conjuntos de datos que exigen procesamiento complejo, bien sea por el volumen de los datos, por su diversidad o porque se generan constantemente. Tales características constituyen un valor agregado y un elemento diferencial que ha permitido que la UCD lidere proyectos de alto impacto con enfoque nacional y territorial.

Al asumir un rol de liderazgo para el sector público en temas de inteligencia artificial, *big data* y aprendizaje de máquina, la UCD también ha dirigido iniciativas para llevar la analítica de datos a otras dependencias y entidades. Por tal motivo, la UCD desempeña además una función transversal ofreciendo el servicio de asesorías técnicas, labor que ha permitido a otras entidades tener un mayor acercamiento a proyectos de analítica de datos y que puedan liderarlos por sí mismas, demostrando que todas cuentan con el potencial para desarrollarlos.

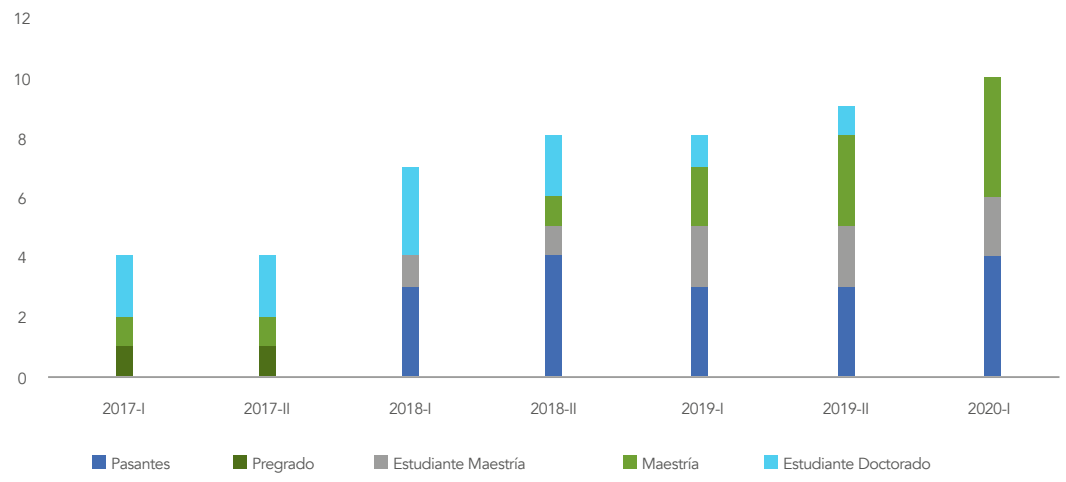
Ejemplos destacados de estas asesorías incluyen el proyecto de análisis de Planes de Gestión Integral de Residuos Sólidos (PGIRS), liderado por la Dirección de Seguimiento y Evaluación de Políticas Públicas (DSEPP); el de análisis automático de PQRSD (peticiones, quejas, reclamos, sugerencias y denuncias), cuyo piloto se adelantó con el Programa Nacional de Servicio

al Ciudadano (PNSC); la evaluación y retroalimentación de proyectos de alto impacto liderados por Alianza CAOBA; la formulación de proyectos para abordar el problema de la accidentalidad vial con la Agencia Nacional de Seguridad Vial (ANSV); y el apoyo a la Dirección de Infraestructura y Energía Sostenible (DIES) en levantamiento del inventario de vías terciarias a partir de imágenes satelitales.

1.3.1.1. Perfiles de los miembros del equipo

Desde la creación de la UCD, el equipo ha contado con personas de diferentes perfiles profesionales, con conocimientos en matemática, estadística, ingeniería y economía. Esta diversidad ha permitido encontrar soluciones desde diferentes perspectivas. En 2018, la UCD comenzó a incorporar estudiantes de últimos semestres interesados en realizar su pasantía o práctica profesional en temas relacionados con analítica de datos. Esa vinculación ha permitido la transferencia de conocimiento y la motivación a nuevas generaciones por especializarse en estos temas. En la figura I.3-1 se muestra cómo ha evolucionado la conformación del equipo de científicos de datos de la UCD.

Figura I.3-1. Perfiles del equipo de la UCD a través del tiempo, 2017-2020 (I semestre)

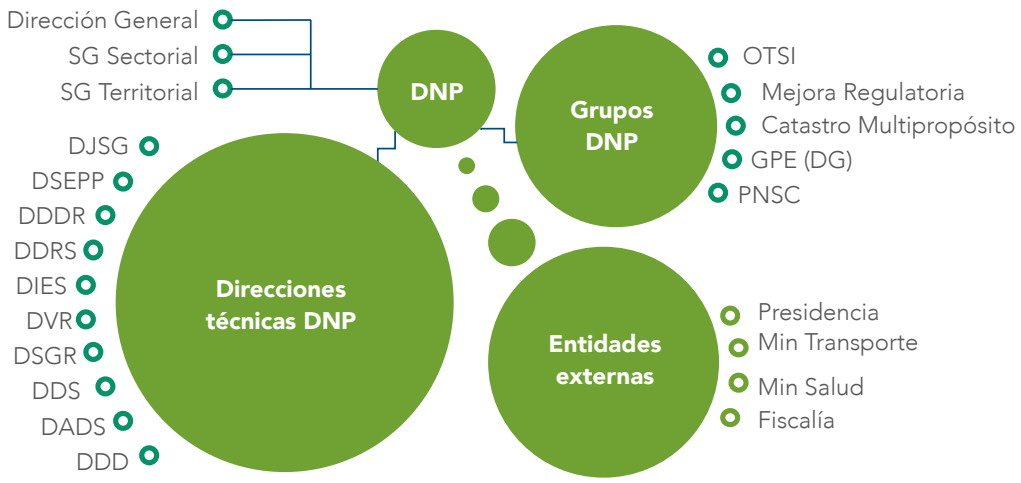


Fuente: elaboración propia.

1.3.1.2. Dependencias y entidades con las cuáles ha trabajado la UCD

La UCD ha trabajado con varias direcciones y dependencias dentro del DNP, también lo ha hecho con diferentes entidades externas con el objetivo de proponer, desarrollar y colaborar en proyectos e iniciativas de analítica de datos que ayuden a la formulación de políticas y creación de valor en el sector público (figura I.3-2).

Figura I.3-2 Dependencias del DNP y otras entidades públicas con las cuáles ha trabajado la UCD



Fuente: elaboración propia.

Además, la UCD ha participado en diversos eventos, compartiendo experiencias y mostrando el impacto de utilizar la explotación de datos para apoyar la formulación de políticas públicas en Colombia. Se destacan como los más relevantes en los que han participado miembros de la UCD siendo ponentes o panelistas los siguientes:

Taller de LatinX in AI Research en la Conferencia ICML **2019** (California)

Seminario Regional sobre Desafíos e Innovaciones en la Medición de la Pobreza y el Seguimiento del ODS-1, organizado por la CEPAL, **2019** (Santiago de Chile)

Analytics Forum, organizado por la Universidad de los Andes **2019 y 2020** (Bogotá)

Congreso de Estadística de la Universidad Nacional, **2018** (Bucaramanga)

Foro Corrupción y *Big Data*, Retos y Oportunidades, **2018** (Bogotá)

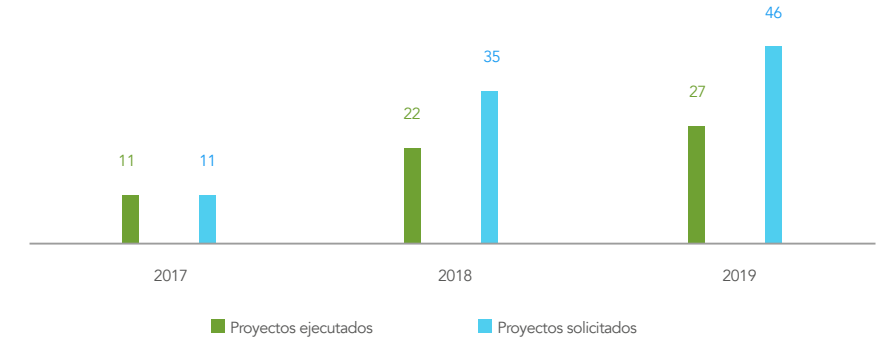
I Foro de Política Nacional de Explotación de Datos (*Big Data*) y ciudades inteligentes, **2018** (Bogotá)

Entre otros eventos.

1.3.2. PROYECTOS ESTRATÉGICOS PARA LA TOMA DE DECISIONES EN EL SECTOR PÚBLICO

Desde el nacimiento de la UCD en 2017, se han desarrollado más de 60 proyectos y pilotos tanto internos del DNP como con otras entidades del Estado, como se ilustra en la figura I.3-3.

Figura I.3-3. Demanda y ejecución anual de proyectos de la UCD



Fuente: elaboración propia.

Debido a la naturaleza sectorial y territorial del DNP, la UCD ha desarrollado proyectos de sectores muy diversos, destacando los proyectos enfocados en la planeación nacional (eje misional del DNP), temas estratégicos para la Administración pública y los relacionados con el sector TIC, como se presenta en la figura I.3-3 (izquierda), donde también se evidencia que muchos de los proyectos comparten un enfoque territorial (derecha).

Figura I.3-4. Número de proyectos desarrollados en la UCD por sector (izquierda) y por enfoque territorial (derecha)

Fuente: elaboración propia.

Además, la UCD ha trabajado con distintos tipos de datos, entre los cuales priman los datos no estructurados, que van desde análisis de texto a análisis de imágenes satelitales. Con respecto a datos estructurados, la complejidad del procesamiento, su volumen o su tasa de generación son algunas de las razones por las que dependencias internas del DNP y las entidades del sector público encuentran un valor agregado en el equipo. En la figura I.3-5 se muestran los proyectos desarrollados por tipo de datos —es posible que un mismo proyecto haya involucrado datos de dos o más tipos distintos—.

Figura I.3-5. Número de proyectos desarrollados en la UCD por tipo de datos

Fuente: elaboración propia.

A continuación, se presentan de manera breve algunos de los proyectos desarrollados por la UCD, junto con los resultados obtenidos. Para conocer más sobre estos y otros proyectos, se puede consultar la sección de la UCD en la página web del DNP, disponible en el siguiente enlace:

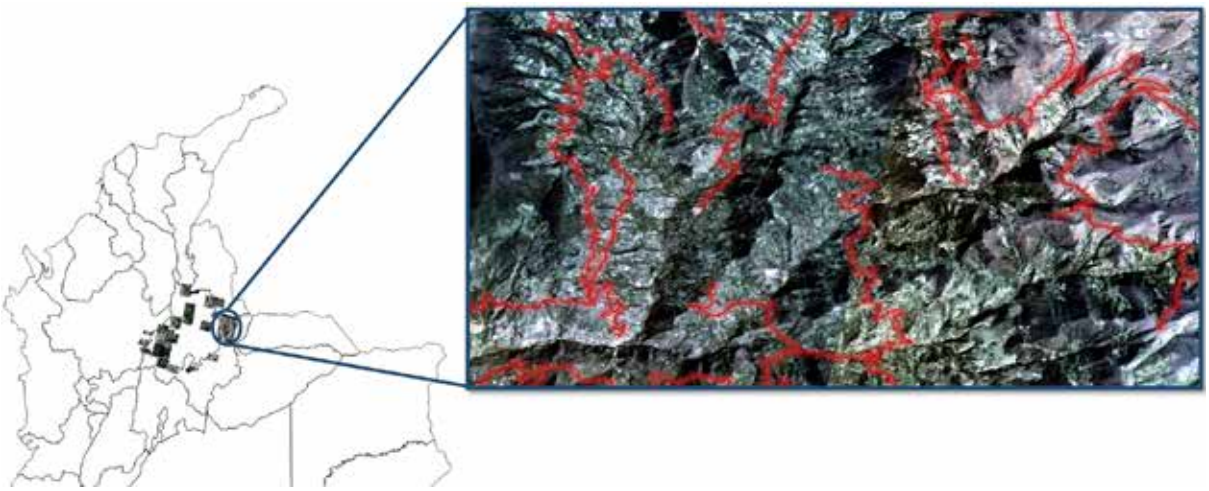
https://www.dnp.gov.co/programas/Desarrollo%20Digital/Paginas/Big%20Data.aspx

1

Identificación de vías terciarias a partir del análisis de imágenes satelitales RGBA

Este proyecto se planteó como un piloto para determinar la viabilidad de construir un inventario de las vías terciarias del país a partir del uso de técnicas de análisis y procesamiento de imágenes satelitales (figura I.3-6); también se incluyó observar la conveniencia de contar con un equipo de geocientíficos para completar el inventario total de las vías terciarias. Para ello se delimitó el ámbito o espacio del piloto de detección de vías en algunas regiones del departamento de Santander, debido a que se dispuso de información de datos georreferenciados de los trayectos de un conjunto de vías terciarias.

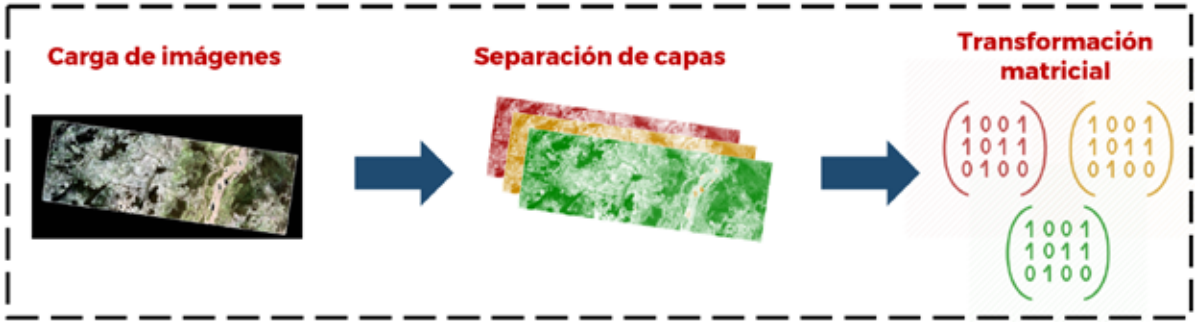
Figura I.3-6. Identificación de vías terciarias a partir del análisis de imágenes satelitales RGBA



Fuente: elaboración propia partir de imágenes del IGAC.

Para cumplir con este objetivo se utilizaron dos algoritmos, *redes neuronales convolucionales y máquinas de soporte vectorial* (*SVM* por sus siglas en inglés), los cuales permitieron obtener resultados preliminares para el análisis de este tipo de imágenes. Para la preparación de los datos fue necesario aplicar una transformación matricial de las imágenes satelitales en cuatro componentes utilizando el modelo de color *"alfa verde azul rojo"*, como se muestra en la figura I.3-7.

Figura I.3-7. Ilustración de transformación matricial de una imagen en matrices RGB

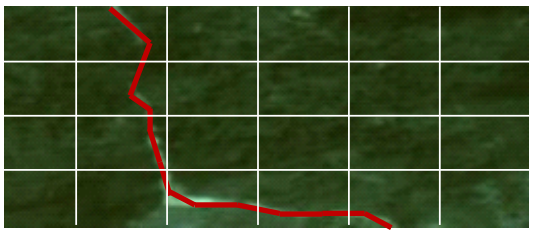


Fuente: elaboración propia.

Después de separar la imagen en capas (matrices) se agrega una capa adicional con la información de la posición de las vías terciarias (*shapefile*) con respecto a las matrices de la imagen; es decir, cada posición (fila, columna) de las matrices de imagen, tendrá asociado un valor adicional de 1 para indicar que por ahí pasa una vía o 0 en el caso contrario. Esa construcción se ilustra en la figura I.3-8. Los datos, con la capa adicional, sirven como insumo para la

etapa de entrenamiento del modelo de SVM y redes neuronales. Posterior a la etapa de entrenamiento, se prueba el modelo con otras imágenes satelitales diferentes a las usadas en el paso anterior, donde el clasificador estimará 1 o 0 para cada pixel de la nueva imagen.

Figura I.3-8. Añadiendo la capa de información georreferenciada de las vías terciarias (derecha) a las capas RGBA de la imagen satelital

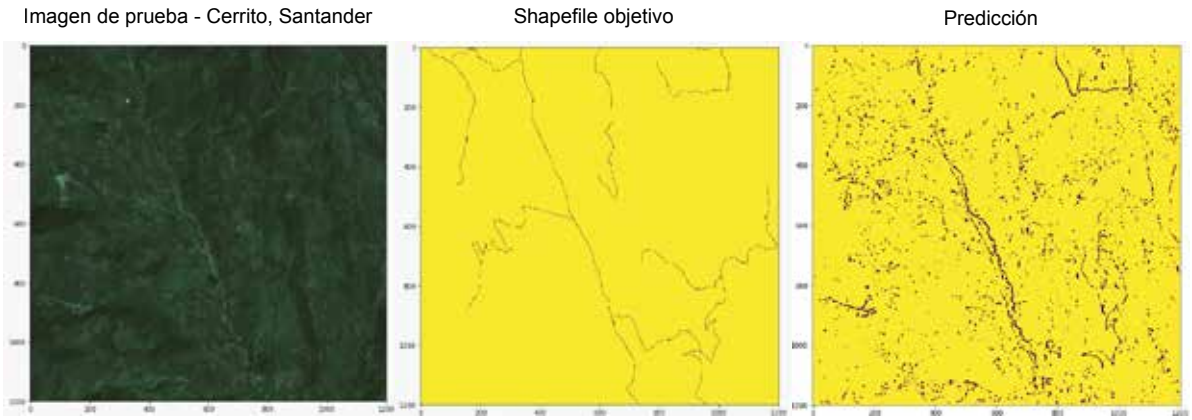


Identificación de vías					
No	Sí	No	No	No	No
No	Sí	No	No	No	No
No	Sí	No	No	No	No
No	No	Sí	Sí	Sí	No

Fuente: elaboración propia a partir de imágenes del IGAC.

Después de observar el porcentaje de acierto de los dos clasificadores probados, SVM (86,5%) y redes neuronales (54%), se opta por el primero. En la figura I.3-9, se muestran los resultados obtenidos para una de las regiones del departamento de Santander, conocida como Cerrito.

Figura I.3-9. Resultados obtenidos para de detección de vías terciarias sobre la imagen satelital de Cerrito (Santander)



Fuente: elaboración propia a partir de imágenes del IGAC.

Los resultados que se obtuvieron en el piloto demuestran la viabilidad para construir el inventario de vías terciarias a partir de metodologías de procesamiento y análisis de imágenes satelitales, también demuestran la necesidad de contar con un grupo de geocientíficos dedicado a la construcción del inventario nacional de vías terciarias.

2

Identificación de ocupación e infraestructura en zonas de ronda hídrica mediante el análisis de imágenes satelitales RGB

Dentro de los planes de ordenamiento territorial (POT) son fundamentales los lineamientos en cuanto a la evaluación y la prevención de riesgos por causas naturales. Específicamente los eventos extremos que pueden causar las crecidas de los ríos, con los que se pone en riesgo la ocupación e infraestructura en la cercanía de los cauces. El artículo 83 del Decreto 2811 de 1974 consagra que debe existir una franja paralela de, como mínimo, 30 metros al cauce de los ríos, denominada *ronda hídrica*.

Con base en esa determinación normativa, la UCD desarrolló una herramienta que utiliza imágenes satelitales RGB —imágenes de tres canales, *Red* (rojo) - *Green* (verde) - *Blue* (azul)— para detectar regiones que presentan ocupación e infraestructura construida dentro de las rondas hídricas. La herramienta combina técnicas de procesamiento de imágenes y técnicas de análisis SIG (Sistemas de Información Geográfica) a fin de generar insumos e información de valor para las comisiones técnicas, para apoyar la toma de decisiones y el planteamiento de recomendaciones con respecto a la delimitación de las rondas hídricas y los planes de ordenamiento territorial.

La herramienta inicialmente descarga información geográfica de OpenStreetMap (OSM); en específico, las capas con datos geográficos de ríos y construcciones. A través de estos y apoyados de técnicas SIG se estima la ronda hídrica de los ríos. Luego se calcula el número de las construcciones que intersecan la ronda, y se obtiene un estimado de la ocupación e infraestructura que puede ser susceptible a inundaciones debido a la crecida del cauce del río, como se ilustra en la figura I.3-10.

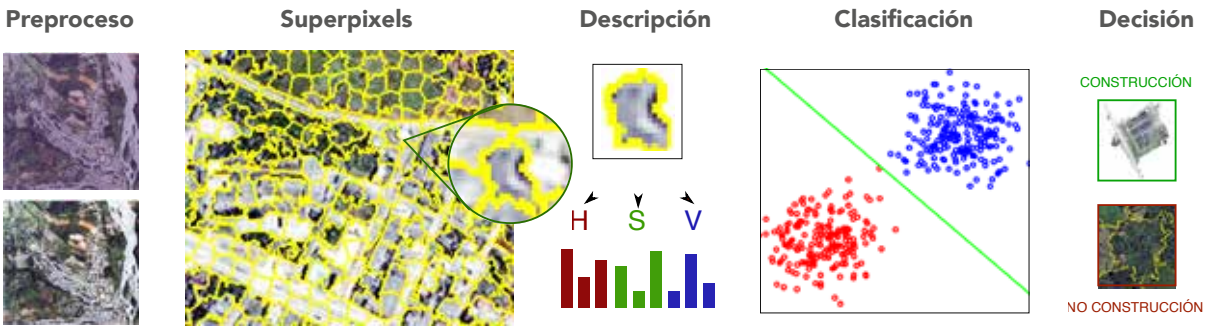
Figura I.3-10. Intersección entre la ronda hídrica de 30 metros (en rojo) y la capa de construcciones de OSM

Nota: Las construcciones en color naranja son susceptibles de inundación por estar dentro de la ronda hídrica.

Fuente: elaboración propia con base en imágenes de OpenStreetMap.

Debido a que la capa de construcciones de OSM solo está disponible para un porcentaje pequeño del territorio nacional, fue necesario estimar las regiones que tienen construcciones mediante el análisis de imágenes satelitales RGB, como se ilustra en la figura I.3-11, obtenidas de los mapas base de Google Maps y ESRI, y transformar esas regiones a información geográfica para hallar las intersecciones con la ronda hídrica estimada a partir de análisis SIG.

Figura I.3-11. Etapas para la detección de construcciones mediante análisis de imagen



Fuente: elaboración propia a partir de imágenes de Google Maps.

Con ello, la herramienta da una visión general del estado de ocupación en las rondas hídricas, provee insumos a las comisiones técnicas municipales y regionales para priorizar la toma de decisiones en cuestión de prevención del riesgo, delimitación de rondas hídricas y planes de ordenamiento territorial. Por ahora la herramienta se encuentra disponible de forma local, para el uso de las direcciones técnicas del DNP.

3 Identificación de proyectos de inversión para el cumplimiento de los ODS

Los Objetivos de Desarrollo Sostenible (ODS) surgen como respuesta a los problemas primordiales de las naciones y buscan optimizar la calidad de vida de sus habitantes. Si bien las metas asociadas a cada ODS contemplan unos indicadores para hacer seguimiento, verificar el avance y su cumplimiento, es necesario contar con indicadores financieros complementarios que permitan elaborar análisis más completos sobre cómo la inversión de recursos públicos se orienta hacia la consecución de cada de esos objetivos. Para ello, la UCD junto con la Dirección de Seguimiento y Evaluación de Políticas Públicas del DNP desarrollaron el proyecto que estima la asignación de recursos de inversión destinados a cada ODS, mediante el análisis y clasificación automática de 112.042 proyectos de inversión financiados con recursos del Sistema General de Regalías (SGR), del Presupuesto General de la Nación (PGN) y recursos de cooperación internacional.

El ejercicio de identificación se efectúa a partir de dos bases de datos: una con los proyectos de inversión por estudiar y otra con textos disponibles en la página de las Naciones Unidas que hacen referencia a los primeros 16 ODS

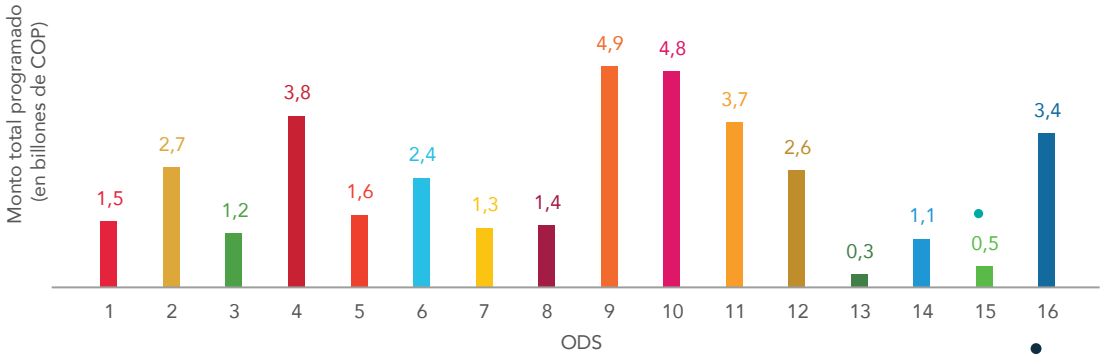
—el ODS 17 se excluye del análisis de texto—. Cada una de las bases de datos se procesó para limpiar el texto y hacer la depuración semántica, de tal manera que se eliminara el texto irrelevante para la clasificación de los proyectos. Luego se usó la metodología TF-IDF (*Term frequency - Inverse document frequency*) para obtener una representación numérica de los textos que permitiera calcular la similitud coseno entre todos los proyectos y cada ODS.

Finalmente, se etiquetaron los proyectos de inversión como pertenecientes, o no, a cada uno de los 16 ODS, se definieron umbrales de clasificación según la distribución de las similitudes calculadas. Este etiquetado múltiple fue necesario porque un mismo proyecto puede contribuir al cumplimiento de más de un ODS, considerando las interrelaciones ampliamente conocidas entre los diferentes Objetivos.
Por ejemplo, un proyecto de soluciones de vivienda (ODS 11) puede contribuir a reducir la pobreza multidimensional (ODS 1) y, al mismo tiempo, impactar en la generación de empleo y crecimiento económico (ODS 8). Este mismo hecho motivó el desarrollo de una herramienta de visualización que brindara una mayor flexibilidad en la clasificación y análisis tanto por un único ODS como de sus interacciones con los demás objetivos.

El principal resultado de la primera versión del proyecto, fue caracterizar el panorama general de la inversión de recursos para la consecución de los ODS, el cual se presentó con detalle en el Reporte Nacional Voluntario de Colombia (2018). Allí se analizó qué ODS muestran mayor relevancia en función de cada fuente de recursos analizada, a la vez que se hace un análisis desagregado por años que permite observar cómo la inversión efectuada en cada ODS ha aumentado o disminuido en el tiempo.

Se encontró así, por ejemplo, que los recursos destinados a cumplir los ODS disminuyeron entre 2015 y 2017 en un 11 %, lo que representó una caída en términos nominales de alrededor de USD 2.850 millones. Igualmente, se identificó una mayor alineación entre algunas fuentes de financiación y ODS específicos; por ejemplo, los recursos del PGN tienen mayor participación en temas de pobreza, salud y educación (ODS 1, 3 y 4), mientras que en los recursos del SGR se destacan temas de infraestructura y construcción de ciudades sostenibles (ODS 9 y 11) y en los de cooperación internacional sobresale la inversión en paz y en acción por el clima (ODS 16 y 13). La figura I.3-12 muestra la inversión identificada con recursos del PGN del 2019 para el cumplimiento de cada uno de los ODS.

Figura I.3-12. Recursos del PGN alineados con cada ODS en 2019



Fuente: elaboración propia a partir de información del PGN.

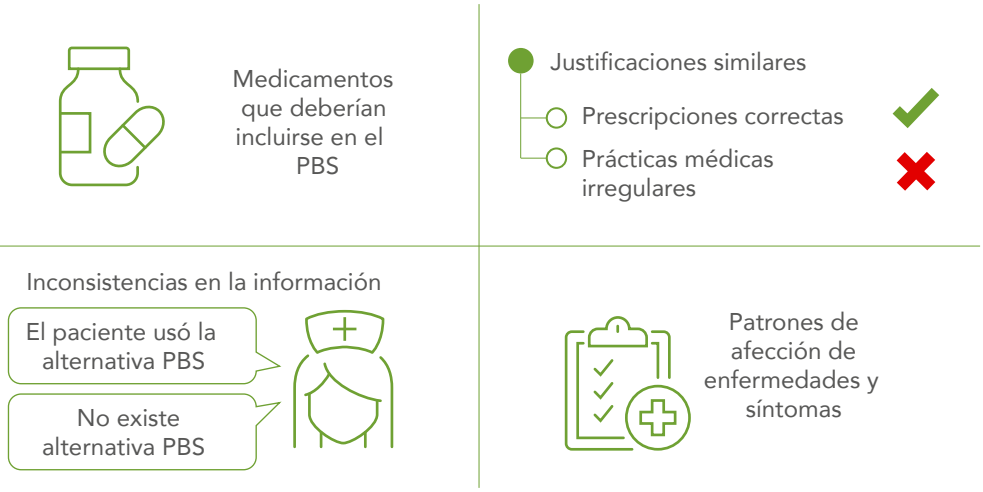
4

Análisis de las justificaciones en prescripciones médicas en la base MIPRES no PBS

MIPRES es una herramienta tecnológica que permite a los profesionales de salud reportar la prescripción de tecnologías en salud no financiadas con recursos de la unidad de pago por capitación (UPC) —dinero que reconoce el sistema a las EPS por la salud de cada afiliado— o servicios complementarios.

El análisis de las justificaciones en las prescripciones médicas adelantado por la UCD utiliza herramientas de análisis y minería de texto para extraer las palabras relevantes que ayudan a relacionar los diagnósticos dados con los medicamentos que se prescriben, con el fin de abarcar los cuatro puntos claves ilustrados en la figura I.3-13. Para ello, se analizaron las justificaciones asociadas a las prescripciones de los medicamentos no incluidos en el Plan de Beneficios en Salud (PBS) registradas en la base de MIPRES en los años 2017 y 2018, equivalente a 5,5 millones de prescripciones, 136 millones de palabras en las justificaciones y 7.500 diagnósticos.

Figura I.3-13. Objetivos del análisis automático de las justificaciones en prescripciones médicas



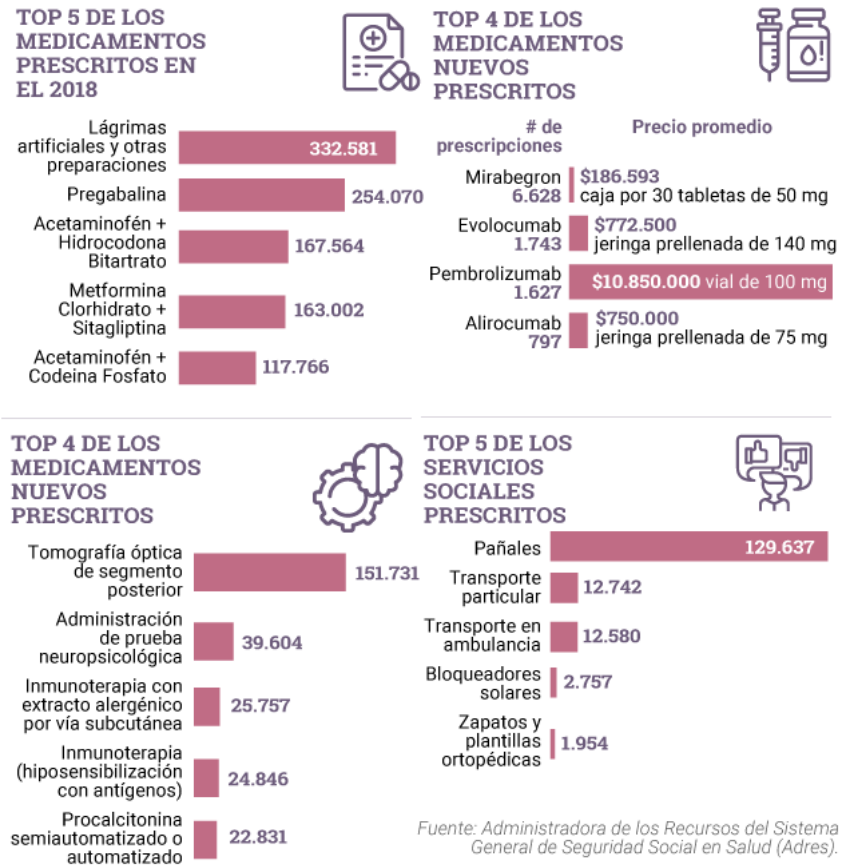
Fuente: elaboración propia.

Las técnicas de análisis y minería de texto permiten revelar relaciones complejas u ocultas dentro de los textos, ello resalta información que no se podría haber encontrado con la lectura de las justificaciones. A partir de los datos mencionados, el análisis de texto genera reportes automáticos, que sirven de insumo a la ADRES (Administradora de los Recursos del Sistema General de Seguridad Social en Salud) y a otras entidades del sector de salud pública para priorizar medicamentos que se deberían incluirse en el PBS —por ejemplo, medicamentos no PBS más prescritos—; la identificación y hallazgo de patrones de prescripciones médicas sesgadas —por ejemplo, irregularidades en la justificación dada por el médico, profesionales que más prescriben tipos de medicamentos no PBS—; validar la información

epidemiológica de otras fuentes y elaborar guías para las buenas prácticas de prescripción médica, inexistentes en la actualidad y tarea pendiente del Documento CONPES 155 de 2012: *Política Farmacéutica Nacional*.

En agosto de 2019, *El Tiempo* divulgó un artículo titulado L*as cinco fórmulas médicas más polémicas que todos pagamos* donde el diario bogotano presenta un análisis de datos estadísticos y relaciones entre medicamentos y diagnósticos que pueden obtenerse con los reportes generados automáticamente como producto de este proyecto. En la figura I.3-14 se muestran algunos de los resultados publicados por la ADRES e incluidos en el artículo de *El Tiempo*.

Figura I.3-14. Datos generados a partir de los reportes automáticos del análisis de las prescripciones médicas de la base de datos MIPRES



Fuente: El Tiempo (2019).

5

Identificación de patrones de reincidencia y reiteración delictiva

En la actualidad, no existe en el país un consenso sobre la definición de reincidencia, ni como cuantificar su ocurrencia. Por otro lado, los datos disponibles por parte de la justicia penal se encuentran desarticulados y no es posible hacer un seguimiento de los casos de reincidencia en el sistema. Este proyecto de analítica de datos permite cuantificar los fenómenos de reincidencia y reiteración en el delito en Colombia, utilizando como fuentes de información las bases de datos del Sistema de Información de Justicia de la Fiscalía (SIJUF) y el Sistema Penal Oral Acusatorio (SPOA). La combinación de estas bases suma más de 40 GB de información, lo que representó un reto al momento de procesar, cruzar y analizar los datos.

A partir del SIJUF y SPOA, se calculó la reincidencia y reiteración con base en dos criterios. El primero, la participación en dos o más casos por parte del indiciado; el segundo, con un criterio construido a partir de los delitos cometidos y de los casos relacionados, teniendo en cuenta la fase del proceso —1: indiciado, 2: imputado, 3: condenado—.

Con base en los análisis efectuados, los resultados muestran que alrededor del 30% de los indiciados, imputados y condenados se clasificaron como reincidentes; además, el 60% de los delitos fueron cometidos por reincidentes como se muestra en las tablas I.3-1 y I.3-2.

Tabla I.3-1. Porcentaje de individuos clasificados como reincidentes

TIPOLOGÍA	NO REINCIDENTES	REINCIDENTES	TOTAL	PROPORCIÓN DE REINCIDENTES
Indiciado	2.914.741	1.451.588	4.366.329	33,2%
Imputado	341.412	149.335	490.747	30,4%
Condenado	244.585	94.632	339.217	27,9%
Total	**3.500.738**	**1.695.555**	**5.196.293**	**32,63%**

Fuente: elaboración propia.

Tabla I.3-2. Porcentaje de delitos cometidos por reincidentes

TIPOLOGÍA	NO REINCIDENTES	REINCIDENTES	TOTAL	PORCENTAJE DE REINCIDENTES
Indiciado	2.914.741	4.230.405	7.145.146	59,2%
Imputado	341.412	612.072	953.484	64,2%
Condenado	244.585	489.597	734.182	66,6%
Total	**350.0738**	**5.332.074**	**8.832.812**	**60,37%**

Fuente: elaboración propia.

6

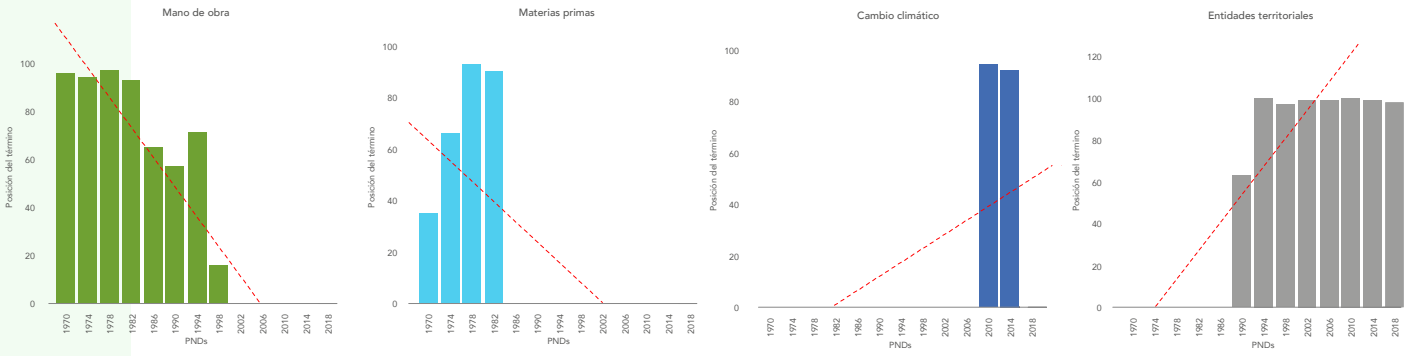
Análisis de los planes nacionales de desarrollo a través del tiempo

El Plan Nacional de Desarrollo (PND) es un documento que contiene la base de las políticas gubernamentales de cada gobierno en Colombia. Por ser la hoja de ruta que establece los objetivos cada administración en el país, fija programas, inversiones y metas para cada cuatrienio. Las bases de cada PND dan, por lo tanto, una buena idea de las prioridades y los temas de interés para cada periodo presidencial.

Con el objetivo de analizar la evolución del contenido y de las temáticas más relevantes tratadas en las bases de los PND a través del tiempo, se utilizaron técnicas de minería y analítica de texto para estudiar los realizados entre 1970 y 2018. El proyecto tenía el propósito de identificar los términos o las palabras más relevantes para ciertos sectores sociales y económicos, y hacer un seguimiento de la frecuencia con la que tales términos se mencionan en los PND con el paso de los años. Un elemento novedoso es que los términos más relevantes por sector no fueron suministrados por expertos temáticos, sino que se obtuvieron mediante técnicas matemáticas, basadas en la comparación de todos los PND con grupos de escritos relacionados con cada sector. Estos documentos, entre los que se encontraban CONPES y otros que contienen políticas públicas, sirvieron como guía para que un algoritmo hallara listas de palabras únicas y las más relevantes para doce sectores, entre los que se encontraban transporte, educación, vivienda, telecomunicaciones y otros.

Con esas listas de términos por sector como insumo, se hizo seguimiento al énfasis que a lo largo de cada PND se ha dado a los distintos sectores y, para cada sector, a la evolución de los términos más representativos a medida que el énfasis cambia. El análisis permite identificar si hay sectores que han ganado o perdido relevancia en los PND con el paso del tiempo; pr ejemplo, la figura I.3-15 muestra el porcentaje de presencia relativa de los términos asociados a cada sector en los PND desde 1990 hasta 2014. La gráfica muestra que el sector de inclusión y reconciliación ganó relevancia durante ese periodo de tiempo, mientras que otros sectores como educación y salud y protección social se han visto ligeramente menos representados en PND más recientes.

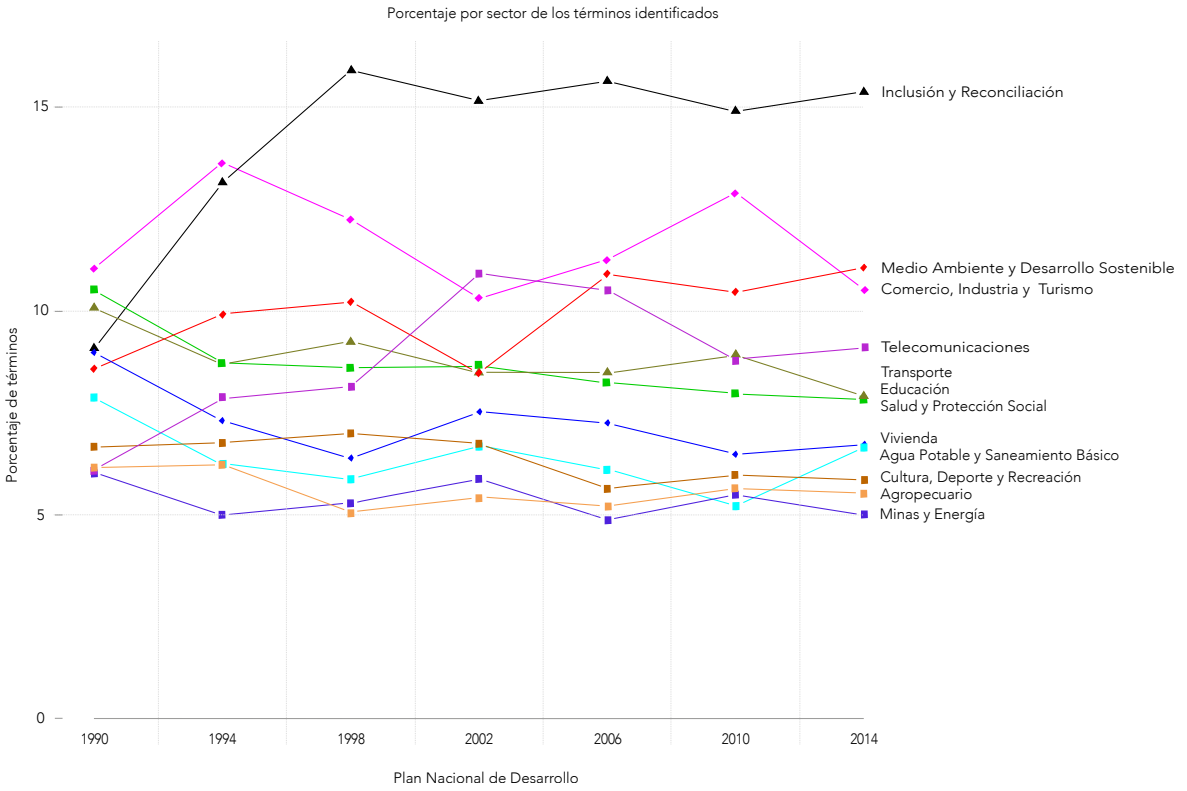
Figura I.3-15. Distribución relativa de los distintos sectores en los PND a lo largo del tiempo



Fuente: elaboración propia a partir de los PND.

Finalmente, para cada uno de los PND entre 1970 y 2018 se hizo un análisis de frecuencias para obtener las 100 palabras y los 50 n-gramas (grupos de 2, 3 y 4 palabras) más repetidos en cada plan de desarrollo. Con la información se elaboraron visualizaciones (nubes de palabras) y análisis de tendencias para las palabras y los bigramas (grupos de dos palabras) más comunes, para estudiar cuáles temas han perdido y ganado relevancia en los PND a través de los años. La figura I.3-16 muestra dos ejemplos bigramas que cada vez aparecen menos en los PND, "mano obra" y "materias primas" y dos ejemplos de bigramas que han ido ganando presencia con el tiempo, "cambio climático" y "entidades territoriales".

Figura I.3-16. Bigramas que tuvieron una pérdida o ganancia significativa de relevancia a lo largo del tiempo



Fuente: elaboración propia.

CAPÍTULO 1.4

REFLEXIONES PARA LA CONSOLIDACIÓN DE LA EXPLOTACIÓN Y ANALÍTICA DE DATOS EN LAS ENTIDADES PÚBLICAS

LA FORMULACIÓN Y EL SEGUIMIENTO DEL DOCUMENTO CONPES 3920, EL TRABAJO LLEVADO A CABO POR LA UNIDAD DE CIENTÍFICOS DE DATOS DEL DNP Y LA TRANSVERSALIDAD DE LA ANALÍTICA EN VARIOS PACTOS DEL PLAN NACIONAL DE DESARROLLO 2018-2022; *PACTO POR COLOMBIA, PACTO POR LA EQUIDAD,* HAN PERMITIDO IDENTIFICAR LECCIONES APRENDIDAS PARA EL APROVECHAMIENTO DE DATOS EN LAS ENTIDADES PÚBLICAS Y LA ELABORACIÓN DE PROYECTOS DE EXPLOTACIÓN DE DATOS.

El planteamiento, desarrollo y despliegue de proyectos de analítica de datos en el sector público para apoyar la toma de decisiones requiere de cinco elementos fundamentales:

1. El fortalecimiento de la cultura de datos en las entidades del Gobierno.
2. Las competencias pertinentes del talento humano.
3. La disposición de datos de calidad.
4. La disponibilidad de herramientas y de tecnología para la explotación de los datos.
5. La consolidación de una gobernanza de datos en las entidades.

Los proyectos de analítica de datos deben proporcionar resultados y herramientas que sean útiles y pertinentes para los tomadores de decisiones en las entidades. Para ello, es fundamental que la analítica de datos esté soportada en una cultura de datos que permita integrar el aprovechamiento de los activos de información con las estrategias institucionales de las entidades públicas. Ello significa que la gestión y el análisis de los datos ha de convertirse en un componente central para tomar decisiones que esté en sintonía con las necesidades y objetivos misionales de cada entidad.

Para lograrlo es necesario que las entidades públicas se aproximen a experiencias nacionales e internacionales que visibilicen el aporte de la analítica de datos para mejorar la gestión pública, focalizar de manera más precisa la intervención de programas y políticas, y mejorar la comprensión de los problemas de carácter público. Ese ejercicio fomenta el interés de los directivos de las entidades del sector público para extraer valor de los datos y, en consecuencia, el diseño y la implementación de acciones y procesos que se requieren para la puesta en marcha tanto de prototipos como de proyectos de análisis de datos.

Una de estas acciones previstas es avanzar en la disponibilidad, la compartición y la apertura de datos de calidad para su explotación. Para ello es fundamental definir una gestión de los datos internos y externos de la entidad durante todo su ciclo de vida, para garantizar que estén disponibles, documentados, seguros, completos y consistentes los más estratégicos para tomar decisiones. La diversidad de las fuentes de datos enriquece los análisis; por tal motivo, es vital que las entidades identifiquen las fuentes de datos útiles para el desarrollo de proyectos de analítica, por ejemplo, datos de registros administrativos, datos abiertos y datos de fuentes del sector privado.

La disponibilidad de datos de calidad para su uso requiere también que las entidades del sector público incorporen mecanismos —tecnológicos, legales y administrativos— para gestionar el almacenamiento y la compartición de datos con otras entidades del Gobierno y organizaciones del sector privado. Para ello es necesario conocer el marco normativo aplicable a la explotación de datos en materia de transparencia y participación ciudadana, protección de datos personales, *habeas data*, interoperabilidad y gestión de archivo documental.

Así mismo, es trascendental conocer y utilizar las guías dispuestas por el Gobierno colombiano en materia de calidad de datos, interoperabilidad, anonimización, registros administrativos y arquitectura TI, que dan lineamientos técnicos generales para que las entidades implementen procesos documentados en la gestión de sus datos.

En relación con la interoperabilidad de datos, se requiere que las entidades conozcan de manera específica los lineamientos técnicos definidos en el Marco de Interoperabilidad del Gobierno Digital colombiano, la disposición normativa y técnica en materia de seguridad y privacidad de la información, de la misma manera que la definición semántica del lenguaje común para el intercambio de información, que habilita la integración y el cambio recíproco de datos entre entidades públicas del país.

Uno de los puntos más fundamentales en la gestión de los datos para la elaboración de proyectos de analítica se relaciona directamente con la incorporación de la ética en el tratamiento de datos y en el desarrollo de algoritmos. En consecuencia, se requiere que las entidades públicas conozcan los principios éticos inherentes al aprovechamiento de datos, e identifiquen buenas prácticas para su gestión. Para ello se debe haber claridad sobre el objetivo público que los proyectos desean alcanzar, al igual que estudiar y analizar los riesgos e impactos que implica nos solo el tratamiento de datos sino también los resultados del proyecto sobre un grupo de valor específico.

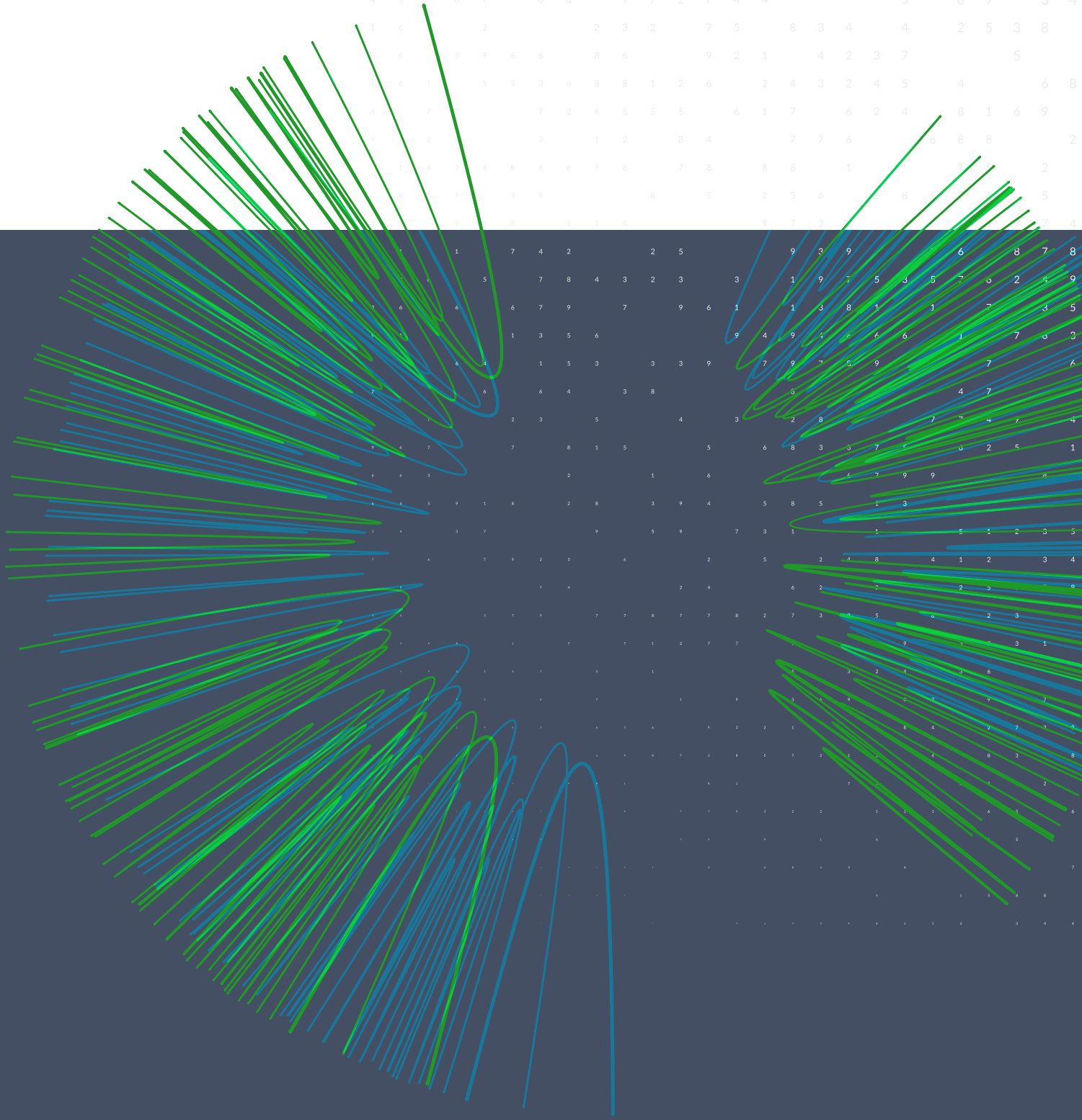
Por otra parte, deben diseñarse estrategias de mitigación para los riesgos enunciados, como la aplicación de técnicas para la anonimización de datos, la disminución de sesgos de clasificación o discriminación, el consentimiento del tratamiento de datos, aunados a la divulgación y comunicación de los resultados del tratamiento de datos para conocimiento de la ciudadanía.

La analítica de datos exige contar con profesionales de perfiles especializados en conocimientos técnicos, pero también en el negocio o sector, para trazar un puente entre las necesidades de la entidad y el análisis de los datos para la posterior toma de decisiones. Uno de los elementos centrales para fortalecer el talento humano del país es el intercambio de experiencias entre grupos de analítica de datos que existen en diferentes entidades públicas.

Esto permite formar una red de conocimiento a través de la transferencia de experiencias exitosas, lecciones aprendidas e intercambio de técnicas, entre otras. Las alianzas entre la académica, el sector privado y las entidades del sector público también son primordiales para fortalecer las habilidades y conocimientos de los funcionarios públicos dedicados a la explotación de los datos.

Con el propósito de aumentar la cultura de datos en las entidades públicas, se recomienda que las mismas participen en espacios colaborativos como el Data Sandbox, diseñado por el Ministerio de las Tecnologías de la Información y las Comunicaciones, cuyo objetivo consiste en fomentar en las entidades el desarrollo de prototipos de analítica de datos y *big data*. El resultado de los prototipos o proyectos piloto en ese tipo de espacios permite evaluar los beneficios en el corto plazo de la explotación de datos; también posibilita identificar las qué requiere cada entidad pública participante para escalar los resultados y alcances de los pilotos.

El aprovechamiento de los datos por parte de las entidades públicas exige la transformación gradual de procesos tecnológicos y organizacionales para enfrentar los desafíos que implica la explotación de los datos para tomar decisiones basadas en la evidencia. La aproximación a los resultados de experiencias nacionales permite identificar lecciones aprendidas y motivar a las entidades públicas de todo el país para construir iniciativas en el corto, mediano y largo plazo que impulsen el aprovechamiento de los datos.



02

PARTE

Guía metodológica para la formulación y ejecución de proyectos de analítica de datos para la toma de decisiones en el sector público

INTRODUCCIÓN

EL DESARROLLO Y PROLIFERACIÓN DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES, JUNTO CON EL AUMENTO DE LA COBERTURA Y ACCESO A INTERNET, HAN PROPICIADO LA GENERACIÓN DE GRANDES VOLÚMENES DE DATOS DE DIVERSAS FUENTES DE INFORMACIÓN, A VELOCIDADES CADA VEZ MAYORES. EN ESTE CONTEXTO, LA CIENCIA DE DATOS SE PRESENTA COMO UNA OPORTUNIDAD PARA MEJORAR LA INNOVACIÓN PÚBLICA A TRAVÉS DE LOS ACTIVOS DE INFORMACIÓN, POTENCIAR LA GENERACIÓN DE VALOR PÚBLICO EN TODOS LOS NIVELES DEL GOBIERNO Y MEJORAR LA TOMA DE DECISIONES EN EL CICLO DE LAS POLÍTICAS PÚBLICAS BASADAS EN LA EVIDENCIA.

El desarrollo e implementación de proyectos de analítica de datos permiten visibilizar la utilidad de los datos para abordar problemáticas de política pública y poner de manifiesto resultados efectivos para atender las necesidades de la ciudadanía. En tal medida, es de interés del Departamento Nacional de Planeación brindar una guía metodológica que oriente a las entidades públicas del orden nacional y territorial en el diseño y desarrollo de proyectos de analítica de datos que soporten no solo la gestión pública sino también la toma de decisiones.

La guía está dirigida a todos los directivos de entidades públicas interesados en incorporar la política de datos como un proceso generador de valor, a los hacedores de políticas públicas basadas en la evidencia y a los funcionarios de las entidades nacionales y territoriales empeñados en incorporar ejes de transformación digital en los planes de desarrollo municipal y departamental.

También está dirigida a líderes de gestión de tecnologías de la información que conozcan la misionalidad de la entidad y los procesos internos relacionados con la gestión y gobierno de datos, a líderes de grupos de información y analítica de datos de las entidades públicas que valoren incorporar mejoras operativas al ciclo de vida de los datos, al igual que a personas vinculadas en grupos de investigación en la academia y el sector privado que reconozcan la trascendencia de generar alianzas con el sector público para fomentar la toma de decisiones basadas en datos.

La guía está compuesta por tres capítulos: en el primero se presenta el inventario de condiciones habilitantes que deben considerar las entidades públicas para iniciar y fortalecer su estrategia de explotación de datos; en el segundo se expone la hoja de ruta para la elaboración de proyectos de analítica de datos; en el tercero se relacionan los actores que participan en el ecosistema de explotación y analítica de datos en el país, y los recursos de los que disponen las entidades públicas para avanzar en una estrategia de aprovechamiento de datos.

CAPÍTULO 2.1

INVENTARIO DE CONDICIONES HABILITANTES PARA LA EXPLOTACIÓN DE DATOS Y EL DISEÑO DE PROYECTOS DE ANALÍTICA

LA FORMULACIÓN Y EL DESARROLLO EFECTIVO DE LOS PROYECTOS DE ANALÍTICA DE DATOS RECLAMA TENER UN PANORAMA GENERAL SOBRE LAS LABORES MISIONALES DE LA ENTIDAD, SU ESTRUCTURA ORGANIZACIONAL Y RELACIONAL CON Y ENTRE LAS DEPENDENCIAS, SUS FUNCIONES Y SINERGIAS. DEMANDA TAMBIÉN DEL CONOCIMIENTO SOBRE LAS CAPACIDADES EXISTENTES EN LA ENTIDAD EN MATERIA DE RECURSO HUMANO, TECNOLÓGICO Y DE LAS CAPACIDADES ESTRATÉGICAS, TÁCTICAS Y OPERATIVAS PARA LA GESTIÓN DEL CICLO DE VIDA DE LOS DATOS.

El enfoque para abordar las capacidades y condiciones habilitantes que se exponen en esta sección, parte de la comprensión del *big data* como un sistema sociotécnico en el que convergen recursos y procesos. En tal sentido, los proyectos de analítica de datos que desarrolle la entidad no serán procesos aislados, sino que estarán integrados y documentados en una estrategia consolidada que responde con los objetivos misionales institucionales. En consecuencia, para responder a este enfoque se requiere un esquema de planeación integral en el que proyectos, prototipos o pruebas de concepto de analítica de datos estén soportados por una gobernanza de datos y por un proceso de gestión del ciclo de vida de los datos.

El fortalecimiento de las capacidades y el aseguramiento de condiciones habilitantes requiere de las entidades de disposición y compromiso por parte, especialmente en relación con la cultura organizacional de la entidad y la transformación de procesos internos. Es importante que las entidades tengan en cuenta que el fortalecimiento de estas capacidades puede incorporarse de manera gradual, de acuerdo con los objetivos estratégicos del proyecto de analítica de datos y en línea con la estrategia de planeación de la entidad.

Esta sección describe las capacidades en términos organizacionales y de recursos que debe implementar una entidad para desplegar una estrategia de explotación de datos. Algunos de los elementos claves que indispensables para la revisión de las capacidades son las personas, la tecnología y los procesos de gobernanza, gestión y operatividad del ciclo de vida de los datos.

Modelo de negocio

La aplicación de este concepto en el ámbito del Gobierno está orientado a definir la estructura organizacional mediante la cual las entidades públicas diseñan e implementan sus procesos y estrategias alrededor de sus objetivos misionales. El modelo de negocio también permite que se integren los recursos invertidos alrededor de uno o varios objetivos estratégicos.

2.1.1. CAPACIDADES EN RECURSOS

Estas capacidades involucran los recursos humanos, tecnológicos y financieros necesarios para abordar una estrategia de explotación de datos en las entidades públicas.

2.1.1.1. Recurso humano

El recurso humano se refiere al talento inexcusable para el desarrollo de proyectos o iniciativas de analítica de datos y para consolidar una estrategia de explotación de datos en la entidad. Este recurso integra los conocimientos y las competencias específicas para la gestión de cada etapa del proyecto de analítica de datos: planteamiento, desarrollo y aprovechamiento. El desarrollo de explotación y analítica de datos es, en esencia, una labor de equipo.

Los equipos de analítica de datos deben estar integrados por perfiles con habilidades y conocimientos matemáticos, analíticos, de programación y en gestión de datos, y un perfil que tenga conocimiento de los objetivos misionales de la entidad y la gestión de las TI.

Este último perfil será el encargado de articular los proyectos de analítica con las características y necesidades de la organización, y con los requerimientos de negocio para responder a las preguntas de interés. Es importante aclarar que el trabajo del equipo de analítica de datos debe estar coordinado con el equipo experto sectorial o experto de políticas públicas con el fin de validar las hipótesis, los resultados y las conclusiones del proyecto que se desarrolle.

El desenvolvimiento de iniciativas de analítica de datos debe involucrar perfiles técnicos para la ingeniería de datos, el modelamiento y el despliegue de las soluciones, el análisis de negocio, y el desarrollo e implementación de infraestructura TI para la integración de soluciones técnicas (figura II.1-1).

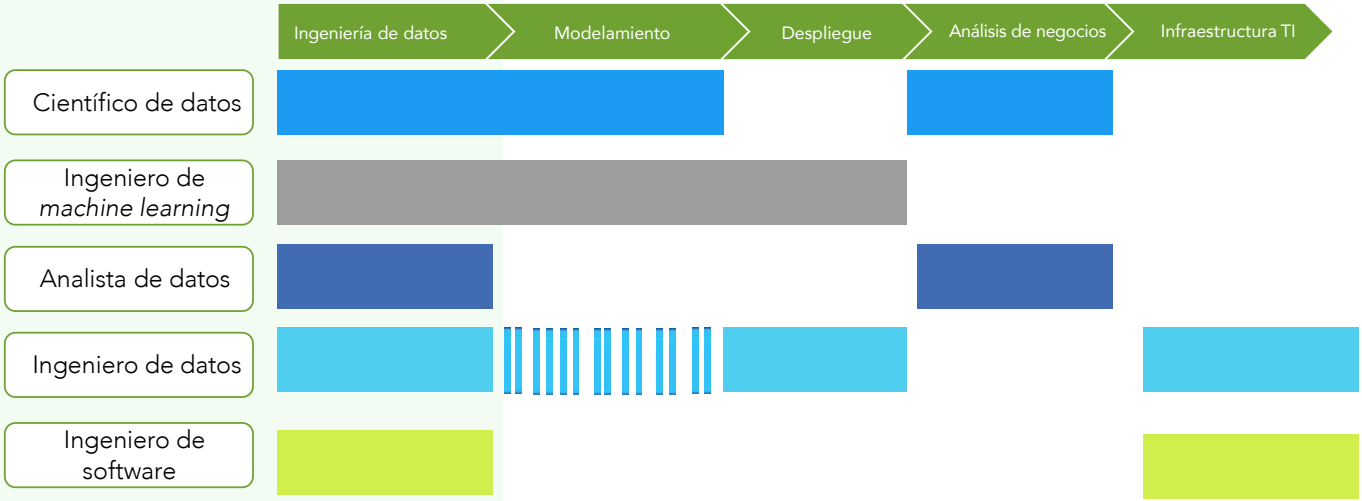
La ingeniería de datos requiere perfiles con conocimientos y habilidades en programación e ingeniería de *software*, y también en *machine learning* y conocimientos en lenguajes de programación como SQL, Python, y Java, y herramientas de *big data* como Hadoop y Hive, entre otros.

El diseño, el entrenamiento y el desarrollo de modelos demanda de habilidades en programación de lenguajes como Python, Julia, Matlab y R, entre otros, y en ocasiones se requiere conocimiento en aprendizaje profundo. Para la etapa de despliegue de modelos y herramientas, los perfiles deben tener conocimientos y habilidades en programación de alto desempeño, idoneidad para escribir código de producción y para la comprensión de tecnologías en la nube.

En cuanto al análisis de negocio, se han de existir perfiles que tengan buenas bases para el análisis y ciencia de datos y a la vez, entiendan el modelo de negocio o el sector de la entidad para poder establecer requerimientos e interpretar resultados. Por último, la infraestructura de TI demanda de perfiles que tengan conocimiento en *software* para escribir código en producción, comprensión del funcionamiento de tecnologías en la nube y, en general, la capacidad para incorporar otros requerimientos de *software* que tengan los productos desarrollados.

Ahora bien, se identifican cinco perfiles pertinentes para la configuración de equipos de analítica de datos (figura II.1-1), que perfiles permiten soportar las necesidades derivadas del diseño, desarrollo e implementación de proyectos de analítica. En los apartados posteriores se sintetizan las aptitudes de cada uno de estos perfiles.

Figura II.1-1. Tipos de perfiles para conformar un equipo de analítica de datos



Fuente: elaboración propia con base en Workera, (pág. 2).

Científico de datos

Su rol principal es definir y llevar a cabo las metodologías apropiadas, tanto de análisis de datos como de desarrollo de modelos estadísticos y de aprendizaje de máquina, de manera que se cumplan los objetivos o se resuelvan las preguntas planteadas en proyecto de analítica. En esa misma medida, debe tener conocimientos de las herramientas de programación y su lenguaje. El científico de datos cumple actividades de ingeniería de datos, modelamiento y análisis del negocio.

Analista de datos

Tiene características similares a los científicos de datos, aunque sus labores están más enfocadas en responder preguntas de negocio mediante consultas, análisis y visualizaciones de los datos disponibles. Este perfil no suele dedicarse al entrenamiento y desarrollo de modelos estadísticos o de aprendizaje de máquina, por lo que no requiere gran fortaleza o experiencia en tales aspectos.

Ingeniero de *Machine Learning*

Este perfil dedica sus esfuerzos durante todo el ciclo de vida de los modelos de aprendizaje de máquina. Esto incluye el procesamiento y adecuación de los datos que alimentarán un modelo, el entrenamiento, ajuste y pruebas de los modelos por desarrollar y su posterior implementación o despliegue, de manera que sus resultados sean aprovechables por los usuarios finales. Este es un rol más especializado que el científico de datos, por lo que se espera que tenga conocimiento amplio de diversas técnicas y arquitecturas de modelamiento, así como de tecnologías y alternativas para el despliegue de modelos.

Ingeniero de *software*

Su función principal es diseñar, implementar y mantener el *software* necesitado para el desarrollo e implementación de soluciones de analítica de datos, así como para la integración de estas soluciones con otros programas o servicios. Para ello, es definitivo tener formación en desarrollo de *software* y conocimientos en lenguajes de programación y desarrollo de aplicaciones que permitan el consumo de datos o el despliegue de aplicaciones de manera eficiente, tanto en tiempo como en recursos de máquina.

Ingeniero de datos

Su rol principal es desarrollar, construir, probar y mantener arquitecturas para el almacenamiento y gestión de los datos, de forma puedan ser aprovechados al máximo de estos. Es el principal responsable por diseñar y establecer el *pipeline*, o flujo de datos, que permite a los demás miembros del equipo tener acceso a los datos que necesitan para llevar a cabo sus labores. Entre sus funciones se encuentran optimizar la recopilación y transformación de datos, optimizar tiempos y recursos y, en general, aplicar soluciones para mejorar la confiabilidad y calidad de los datos. Exige conocimientos en diseño de bases de datos y sistemas de almacenamiento SQL y NoSQL, técnicas como MapReduce y tecnologías y lenguajes enfocados a *big data*. Este perfil se puede ver como una combinación entre el ingeniero de *software* y el científico de datos.

Configurar equipos que reúnan los perfiles y las competencias descritas garantiza que la entidad diseñe e implemente adecuadamente proyectos de analítica de datos. Las habilidades aunadas a los conocimientos de los integrantes serán útiles para generar nexos de comunicación con el nivel directivo e identificar necesidades de la entidad, a fin de diseñar proyectos y estrategias de explotación de datos que respondan con los requerimientos y objetivos propios de la organización.

Más allá de las habilidades técnicas (o "duras") indispensables, las personas que forman parte del equipo de analítica de datos deben tener conocimientos sobre las implicaciones éticas y legales del manejo de datos, la privacidad y la protección de datos personales, al igual que de la propiedad intelectual. Sus habilidades deben responder a la aplicación de soluciones tecnológicas, procedimientos y técnicas para proteger la privacidad de la información, como, las de anonimización de datos personales.

Otra de las competencias vitales en las personas que integran estos equipos, es la capacidad de visualizar y comunicar a los tomadores de decisiones los datos de manera acertada, visualmente llamativa, simple y oportuna, dado que el objetivo principal de la analítica de datos es transformarlos en información útil, clara y pertinente para los fines determinados.

También es preponderante considerar la existencia de competencias de comunicación y comprensión de problemas en el momento de estructurar equipos de analítica de datos, pues estas afectan de manera positiva tanto la calidad como los resultados de los proyectos de explotación de datos. Entre estas competencias se encuentran dos capacidades primordiales; la primera, para entender necesidades de negocio y luego traducirlas en proyectos de analítica; y la segunda, para comunicar y presentar sus proyectos y resultados a personas con diferentes niveles de conocimiento en estos temas, de manera que ellas los comprendan y puedan ser aprovechados por el resto de la entidad.

PREGUNTAS ORIENTADORAS

- ¿Existe dentro de la entidad un equipo encargado de gestionar, procesar, analizar, interpretar y almacenar los datos?
- ¿Hay funcionarios o contratistas en la entidad que hayan recibido capacitación o entrenamiento en analítica de datos?
- ¿La experiencia del recurso humano con el que cuenta la entidad permite tomar decisiones basadas en datos?
- ¿La entidad cuenta con perfiles que gestionen datos y a su vez tengan una comprensión de las principales necesidades del negocio de la entidad?
- ¿El equipo de tecnologías de la información o soporte técnico de su entidad tiene conocimiento para establecer una infraestructura tecnológica robusta que permita gestionar el ciclo de vida de los datos y desarrollar proyectos de analítica?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
Invertir en un equipo de analítica de datos muy especializado sin tener claras las necesidades de la entidad.	Hacer un estudio de necesidades que tiene la entidad en temas de analítica de datos, para identificar cuáles técnicas se requieren aplicar y, con base en eso, vincular los perfiles. Muchas veces la analítica avanzada no es la mejor solución a un problema que se puede resolver de manera más sencilla.
	Buscar perfiles específicos de acuerdo con las necesidades del negocio y las habilidades que se requieren para el desarrollo de proyectos de analítica de datos.
	Apoyarse en experiencias de otras entidades públicas para identificar la pertinencia de vincular ciertos perfiles, aplicar técnicas, y estrategias de analítica de datos.
Dificultades para satisfacer la búsqueda de perfiles por ser demasiado específicos dadas las necesidades de la entidad.	Vincular personas con capacidades básicas en analítica y capacitarlas de acuerdo con las necesidades que se requieren frente al modelo de negocio de la entidad.
	Capacitar y formar el recurso humano que ya existe en la entidad, teniendo en cuenta que el desarrollo de competencias es un proceso de aprendizaje continuo que demanda permanentemente sesiones capacitaciones y trabajo colaborativo.
Invertir dinero y esfuerzos en adquirir perfiles muy técnicos para la explotación de datos sin haber definido previamente políticas de gobernanza de datos y de gestión de los datos durante todo su ciclo de vida.	Definir, de manera previa, un esquema de gobernanza de datos y gestión de datos, donde se determinen de manera estandarizada todos los protocolos, las normativas, las políticas y los procesos para gestionar el ciclo de vida de los datos, y se atiendan asuntos relacionados con la calidad de los datos, la seguridad y privacidad de la información, la recolección, el almacenamiento y el flujo de la información. Es importante que el equipo de analítica de datos sepa con qué datos, tanto internos como externos, cuenta la entidad, al igual que los protocolos y procesos para acceder y tratar los datos, entre otros asuntos.
Alta rotación de personal especializado en explotación y analítica de datos, principalmente por sobredemanda en el mercado laboral para este tipo de perfiles.	Este es un riesgo externo, en el que la entidad pública tiene poco control. Sin embargo, se recomienda que la entidad fomente planes de formación, experiencias académicas y profesionales nacionales e internacionales que motiven a los perfiles a continuar en la entidad.
	Es importante que la entidad tenga procesos de conservación y transferencia de conocimiento de forma que, ante una eventual rotación de personal, la capacitación de las nuevas personas sea más rápida y menos costosa.
	Es importante documentar y divulgar el desarrollo de proyectos de analítica de datos para conservar la trazabilidad de los proyectos que se han elaborado en la entidad.
Falta de alineación entre el equipo conocedor del negocio y el equipo técnico para el entendimiento común de las preguntas de negocio.	Es importante tener espacios de conversación inicial, donde queden claros los objetivos por alcanzar a través de la analítica de datos. También cabe hacer reuniones de seguimiento a fin de validar que el mensaje del equipo de negocio se esté entendiendo de manera adecuada por parte del equipo técnico.

2.1.1.2. Recursos tecnológicos

El recurso tecnológico hace referencia a la infraestructura (*hardware* y *software*) que permite extraer el valor a los datos. Para el desarrollo de proyectos de analítica de datos es vital contar infraestructura robusta capaz de recopilar datos de distintas fuentes de información, almacenarlos, procesarlos, consultarlos, visualizarlos y utilizarlos como insumo para otros procesos. Para lo anterior, se requiere disponer de un entorno de trabajo que administre, distribuya, controle y procese los datos de los sistemas de computación y almacenamiento. Los esquemas de infraestructura que más son empleados por las organizaciones pertenecen a los siguientes tres tipos:

1

En la infraestructura *On-premise*

Las bases de datos y las herramientas para su explotación se encuentran disponibles en los servidores de la entidad. En este tipo de modalidad, la seguridad y gestión de la infraestructura están a cargo de la entidad y la inversión inicial es mucho mayor, ya que es prioritario adquirir servidores, centros de datos y sistemas de seguridad de la información. Las inversiones relacionadas con este tipo de infraestructura son de tipo CAPEX.

2

En la de tipo "nube" pública o privada

Las bases de datos y las herramientas para su explotación se encuentran disponibles en los servidores de terceros, a los cuales se accede a través de Internet. En este tipo de solución, la entidad comparte el servidor ofrecido por el proveedor con otras empresas y entidades. Únicamente se requiere el acceso a una aplicación, mientras que el servicio en la web presta el alojamiento de los recursos y programas para la gestión de los datos. En este tipo de modalidad, la administración de la infraestructura se encuentra a cargo del proveedor del servicio en la nube.

La modalidad anterior permite *acceder a Infraestructura como servicio (IaaS) Plataforma como servicio (PaaS) y Software como servicio (SaaS).* En la infraestructura como servicio, el proveedor ofrece el espacio de servidores, almacenamiento, red, conexión a Internet, y el monto de las inversiones se reduce considerablemente, pues el servicio se presta conforme a las necesidades de la empresa o entidad. En las plataformas como servicio, el proveedor ofrece la posibilidad de desarrollar y adquirir aplicaciones con una función específica —servidor de correo, servidor de servidores— y presta la infraestructura esencial para ello. Entre las principales ventajas que se identifican está la disponibilidad de una plataforma de pruebas y el desarrollo para alojar las aplicaciones en un único entorno, y la facilidad de realizar trabajo colaborativo. En el *software* como servicio los clientes o usuarios se conectan al *software* con un propósito específico a través de un API (*Application Programing Interface* por sus siglas en inglés) del proveedor o por medio de la web. Una de las principales ventajas que se identifica en este caso es que los usuarios no tienen la necesidad de gestionar el *software*, pues este servicio lo presta el proveedor.

3

La infraestructura Híbrida

Corresponde a la compuesta por las dos modalidades descritas anteriormente. En este caso, la solución tecnológica se encuentra disponible en el servidor de un proveedor; sin embargo, los datos del tipo más privado estarían alojados en servidores de la entidad. Para ello, se requiere la integración total entre la infraestructura interna de la entidad y los servicios provistos en la nube, con un único entorno de control para su gestión. Bajo esta modalidad, la entidad mantiene el control de la infraestructura interna de bases de datos y de servidores y sube a la nube únicamente lo que crea pertinente de acuerdo con criterios de simplificación, flexibilidad y bajo costo.

La infraestructura TI de cada entidad debe tener un diseño que le permita ser interoperable con sistemas de información de otras entidades públicas, con el fin de facilitar el intercambio de información entre instituciones. El marco de interoperabilidad del Gobierno nacional[6], definido en el año 2019 por el Ministerio de las Tecnologías de la Información y las Comunicaciones (MinTIC), establece la condición de que las entidades públicas diseñen una arquitectura de infraestructura tecnológica que esté en sintonía con las necesidades de intercambio de información.
Esa relación de aspectos técnicos involucra especificaciones de la interfaz, los protocolos de interconexión, los servicios para la integración de los datos, los protocolos de comunicación seguros, servidores de seguridad y las soluciones integradas para la interoperabilidad, entre otros.

El desarrollo efectivo y pertinente de la infraestructura de datos para *big data* y analítica de datos de una entidad, debe estar totalmente relacionado con la visión integral de la gestión TI, en la que se incluye una planeación estratégica de la infraestructura tecnológica en un horizonte de corto, mediano y largo plazo. En ese sentido, tanto las inversiones como el desarrollo tecnológico de la entidad han de estar soportados en las recomendaciones y los lineamientos del Ministerio de las Tecnologías de la Información y las Comunicaciones, los cuales acogen el Marco de Referencia de Arquitectura Empresarial para la Gestión TI[7].
Estos lineamientos van en línea con los objetivos estratégicos para la transformación digital del país, para la adopción de tendencias tecnológicas de la cuarta revolución industrial como el *big data* y para la toma de decisiones basadas en datos. Con este marco de referencia, las entidades disponen de herramientas para la construcción de los planes estratégicos de las tecnologías de la información (PETI) y la consolidación de la Arquitectura TI, entre otros.

Es significativo resaltar que en el articulado del Plan Nacional de Desarrollo 2018-2022: *Pacto por Colombia, pacto por la equidad*, se definió en el Pacto VII: *Pacto por la transformación digital del país*, dos de los doce principios para el desarrollo de proyectos estratégicos de transformación digital, a saber: 1) la priorización de los servicios de nube, con el fin de optimizar la gestión de los recursos públicos en proyectos de TIC, y 2) la promoción de tecnologías basadas en *software* y código abierto.

6. Disponible en: http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_ gobierno_digital.pdf
7. Disponible en https://www.mintic.gov.co/arquitecturati/630/w3-article-9440.html

En la tabla II.1-1 se presenta un resumen de las ventajas y desventajas principales de cada tipo de infraestructura.

Tabla II.1-1. Características por tipo de infraestructura

TIPO DE INFRAESTRUCTURA	COSTOS	ESCALABILIDAD	SEGURIDAD
On-premise	Los gastos de capital inicial son mayores dada la necesidad de adquirir y mantener servidores, centros de datos y sistemas de seguridad de la información de acuerdo con una necesidad.	Es más rígida la posibilidad de hacer actualizaciones e implementaciones.	• Control directo de a configuración, manejo y seguridad de los datos. • Necesidad de tener un esquema de seguridad de sistemas y datos más robusto.
Nube	Los gastos de capital son bajos al inicio y más graduales dado que depende de la escalabilidad de las necesidades.	• Rapidez y control para implementar actualizaciones. • Adquisición de servicios, plataformas y *software* escalable de acuerdo con la necesidad.	La seguridad está a cargo del proveedor de la nube.

Fuente: elaboración propia.

PREGUNTAS ORIENTADORAS

- ¿La entidad cuenta con servidores, capacidades de almacenamiento, memoria, servicios en la nube y equipos informáticos para la analítica de datos?
- ¿En la entidad se están implementando soluciones tecnológicas orientadas a la analítica de datos?
- ¿La entidad cuenta con fuentes de datos no relacionales, semiestructuradas o no estructuradas?
- ¿La entidad conoce de los requisitos técnicos para interoperar servicios digitales?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
Invertir en *hardware* y *software* que no responda a las necesidades de la entidad, decisión que, en algunos casos, deriva en gastos innecesarios o fallas técnicas.	Considerar la adopción de soluciones tecnológicas con carácter demostrativo y de prueba para iniciativas en explotación de datos.
	Considerar explorar la adquisición de servicios en la nube para suplir las demandas de explotación de datos sin necesidad de almacenamiento en servidores o alojamientos propios.
	Considerar el uso de *software* y marcos de código abierto como solución tecnológica a los desafíos de la analítica de datos. Para ello se recomienda analizar criterios de decisión para la adquisición tecnológica como estabilidad, actualizaciones frecuentes, soporte técnico necesario.
No tener claridad sobre las capacidades de almacenamiento y procesamiento que requiere la entidad en el corto y mediano plazo.	Realizar una consultoría o estudio que identifique las necesidades que tiene la entidad en materia de almacenamiento y procesamiento de datos incorporando un enfoque prospectivo.
	Con base en el estudio anterior, adquirir una infraestructura en nube pública o privada que permita su escalamiento gradual conforme a las necesidades institucionales en un horizonte de corto y mediano plazo.
Fallas en la seguridad y privacidad de la información por falta de medidas de mitigación de riesgos.	Si la infraestructura es *on-premise* es necesario tener personal especializado en seguridad que haga un monitoreo constante de los sistemas para identificar riesgos y necesidades de mejora. Si es infraestructura en nube, es necesario asegurar el acceso a los recursos, y tener herramientas legales o copias de seguridad contempladas en el caso de que el proveedor falle. En ambos casos es importante considerar los lineamientos del MinTIC sobre la gestión de la seguridad y privacidad de la información y los lineamientos del uso de nube pública.

2.1.1.3. Recursos financieros

Los recursos financieros comprenden la disposición y la capacidad de la entidad para orientar medios monetarios al despliegue de una estrategia de explotación de datos y el desarrollo de proyectos de analítica. Sobre la base de la planeación financiera es posible la adquisición del talento humano e infraestructura tecnológica, que soporte el desarrollo de proyectos de analítica de datos. El monto de inversión destinada a proyectos de explotación de datos depende del esquema de planeación de la entidad y de la priorización de rubros para el aprovechamiento de datos; por ende, lo más recomendable es que el esquema de financiación sea escalable, es decir que en los proyectos se delimite el alcance, las metas y los objetivos.

El cálculo o la estimación del retorno de la inversión financiera en proyectos de analítica de datos aún no está estandarizado, además, es especialmente complejo en el sector gobierno, donde los objetivos de cualquier proyecto o estrategia responden especialmente a la generación de beneficio social y de valor público. Sin embargo, se recomienda que la entidad identifique en su plan de inversión en explotación de datos cuál es el monto de la inversión en capital humano y en infraestructura tecnológica que demanda el proyecto; igualmente ha de identificar los beneficios en términos de eficiencia y efectividad que podría alcanzar la entidad a partir de este. Una aproximación puede ser a partir de ahorros en costos de operación o beneficios sociales de los ciudadanos.

PREGUNTAS ORIENTADORAS

- ¿La entidad ha realizado una estimación de las necesidades tecnológicas y en recurso humano que requiere para desarrollar una estrategia de analítica de datos?
- ¿La entidad ha destinado un porcentaje de su presupuesto para actividades de analítica y explotación de datos?
- ¿La entidad incorpora como un hito de la planeación, la destinación de recursos financieros para el mejoramiento de las capacidades institucionales para avanzar en una estrategia de explotación de datos?
- ¿Se ha estimado de manera previa cuál podría ser el aporte de un proyecto de analítica de datos para generar valor dentro de la entidad?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
Invertir recursos financieros en una estrategia de explotación de datos que no responda con las necesidades de la entidad.	La focalización de recursos debe ir alineada con un plan que integre los recursos y capacidades necesarios para el despliegue de una estrategia de explotación de datos en la entidad.
Invertir recursos en proyectos muy ambiciosos que no tengan metas y objetivos cumplibles.	Iniciar con pruebas de concepto cuyos resultados, además de medibles y cuantificables, puedan escalarse en el mediano plazo.
Invertir recursos en una estrategia de explotación de datos sin antes determinar cuál es el impacto o beneficio en términos de valor público.	Identificar cuál es el valor público que genera el desarrollo del proyecto de analítica de datos, que no necesariamente tiene que ser una estimación cuantitativa. Como primera medida es necesario reconocer cuál es el beneficio que le representa a la entidad y a los ciudadanos la inversión en explotación de datos.
Invertir recursos financieros para la contratación de talento humano antes de considerar alianzas con otros actores, para vincular o capacitar talento humano.	Considerar alianzas o esquemas de colaboración con la academia como pasantías o ejecución de proyectos de grado para vincular personal que apoye al equipo dentro de la entidad. Gestionar alianzas con el sector privado y con el MinTIC para vincular personas que apoyen los proyectos de analítica, a través de iniciativas o programas de talento en analítica de datos.

2.1.2. CAPACIDADES ORGANIZACIONALES

Las capacidades organizacionales permiten establecer las bases internas la entidad para extraer el valor de los datos durante todo su ciclo de vida. Esto, a su vez, facilita la disponibilidad de datos de calidad para el desarrollo de proyectos de analítica, en cumplimiento con la normativa de privacidad y seguridad de la información. La planeación estratégica, táctica y operativa para el uso de los datos soporta el desarrollo de proyectos de analítica y permite orientar de manera eficiente la utilización de los recursos, mencionados en la sección II. 1.1, hacia los objetivos misionales y el modelo de negocio de la entidad.

Los tres niveles de la planeación —estratégico, táctico y operativo—, incorporan temas comunes relacionados con la captura y generación de datos, almacenamiento y documentación de los datos (metadatos), gestión de datos maestros, gestión de arquitectura de datos, gestión documental de la entidad, gestión de la calidad de los datos y de protección y privacidad de la información. Sin embargo, como se mostrará más adelante, los tres niveles tienen características particulares, así: el *nivel estratégico* involucra a instancias directivas de la entidad, asigna roles y responsables para la gestión del ciclo de vida de los datos y estructura tanto el esquema normativo como de política acorde con los objetivos misionales de la entidad. El *nivel táctico* traza las tareas, las estrategias, los protocolos y las prácticas que materializan la gobernanza de datos definida en la planeación estratégica. Por últimos, el *nivel operativo* corresponde a la ejecución de las tareas y prácticas definidas previamente en el nivel táctico.

2.1.2.1. Capacidad estratégica

La gobernanza de datos es un conjunto de políticas, normas y métricas que permiten integrar los roles y procesos relacionados con la gestión de los datos. A partir de la gobernanza de datos, se establece una estrategia integral y estandarizada del uso de datos en la entidad, de tal manera que estén disponibles, organizados, documentos y seguros. Además, la gobernanza de datos permite alinear los recursos de la entidad con el modelo de negocio y planificar y coordinar los procesos de levantamiento, flujo y uso de datos en la entidad para la consecución de sus objetivos misionales.

La gobernanza de datos involucra una dimensión política estrechamente relacionada con los procesos de coordinación desde los cargos directivos y de alta gerencia, sobre las medidas y lineamientos para el desarrollo de los ecosistemas de datos y para el uso y explotación de *big data*. Los procesos de política de datos otorgan lineamientos sobre la gestión de datos de la entidad y esquemas de seguimiento y evaluación que involucran la definición de indicadores y metas para tomar decisiones. Al mismo tiempo, la política de datos coordina el despliegue generalizado del ecosistema de datos, direccionando las áreas, las dependencias y los equipos de trabajo de la entidad que realizan actividades relacionadas con la gestión, analítica y explotación de datos.

La gobernanza de datos también involucra la definición de un marco normativo para la gestión, la protección, el almacenamiento y la explotación de los datos. Con esas directrices se definen protocolos para garantizar el cumplimiento de privacidad de datos personales, a través de lineamientos que regulen tano el uso como la divulgación de los datos. También se tiene en cuenta en este aspecto la definición de un marco ético para la gestión de los datos a lo largo de todo su ciclo de vida. Dentro de marco enunciado se definen principios éticos para su uso y tratamiento —por ejemplo, principio de finalidad, de no discriminación, de seguridad, de acceso de la persona interesada, entre otros—. También se establecen protocolos con los cuales anticipar los riesgos generados por la explotación de datos —privacidad, discriminación algorítmica, opacidad—, y planificar acciones para mitigarlos —por ejemplo, la privacidad y ética desde el diseño de cada una de las etapas de los proyectos de analítica de datos—.

PREGUNTAS ORIENTADORAS

- ¿La entidad tiene identificado cuáles son los objetivos estratégicos que se quisieran alcanzar a partir del desarrollo de un proyecto de analítica de datos?
- ¿Cómo es la cultura de la entidad en torno a la toma de decisiones basadas en datos?
- ¿La entidad ha realizado un análisis técnico sobre las oportunidades de la implementación de un modelo de explotación de datos y *big data*?
- ¿Existe un manual interno que reglamente el uso y la compartición de datos entre las dependencias de la entidad?
- ¿La entidad tiene definida una política de gobernanza de datos que incorpore los elementos de privacidad, estándares de datos, archivo y preservación de los datos, y reúso?
- ¿Cómo es la gestión de metadatos de la entidad?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
El proyecto de explotación de datos no tiene el apoyo directivo de la entidad y no genera incidencia.	Fortalecer la comunicación entre el equipo de analítica de datos y los tomadores de decisiones en la entidad, es decisivo para aprovechar el proyecto y garantizar su posterior utilización.
	Incorporar la cultura de datos como parte de la estrategia organizacional que permita superar la toma de decisiones de forma intuitiva y sin evidencia.
Los conjuntos de datos no tienen una propiedad clara y no se tiene registro de los responsables de esos datos.	Incorporar en la estrategia de gobernanza de datos, roles y responsabilidades para la gestión de los activos de información de la entidad. La responsabilidad sobre los conjuntos de datos debe estar definida y visible.
Una vez implementado el proyecto de analítica de datos, o logrado un avance en la estrategia de explotación de datos, no hay claridad frente al impacto generado.	Definir desde el diseño del proyecto o la estrategia de explotación de datos cuáles son los indicadores o métricas mediante los cuales se medirá el resultado del proyecto. Estos deben servir para mejorar la gobernanza de datos de la entidad.
La entidad no conoce cuáles son los datos más estratégicos para su organización y, en esa medida, tampoco los gestiona ni genera valor a partir de ellos.	Identificar y definir los conjuntos de datos estratégicos para la entidad. La gestión de estos datos permitirá mantenerlos actualizados y administrados de acuerdo con los roles y responsabilidades definidos en la gobernanza de datos.

2.1.2.2. Capacidad táctica - Gestión del ciclo de vida de los datos

La capacidad táctica articula los criterios de gobernanza definida en la dimensión estratégica con los procesos operativos que permiten la explotación de datos. En esta dimensión es importante definir el "cómo" se llevarán a cabo estos procesos. Dentro de la capacidad operativa se contemplan dos actividades importantes para la adecuada implementación de la gobernanza de datos definida con anterioridad: la gestión de proyectos y la gestión de los datos.

2.1.2.2.1 Gestión de proyectos

Es el conjunto de metodologías requeridas y pertinentes para la dirección y desarrollo de proyectos de analítica de datos. Con esta base se espera que la entidad pueda gestionar los proyectos de analítica de manera progresiva y esquematizada, en línea con los estándares definidos en la gobernanza de datos. La gestión debe estar relacionada con el ciclo de planeación de proyectos, con ellas se han de adoptar mecanismos de seguimiento y evaluación de sus resultados y estrategias de gestión del conocimiento, los cuales deben aplicarse incluso en proyectos piloto y pruebas de concepto.

2.1.2.2.2 Gestión de datos

La gestión de los datos relaciona todos los procesos sistemáticos y estandarizados que permiten satisfacer las necesidades del ciclo de vida de los datos para crear valor a partir de ellos. Este ciclo está empalmado con todas las prácticas, conceptos y protocolos para que la entidad pueda mantener el control de datos externos e internos, y los procesos para garantizar calidad, seguridad, integración y gestión de los datos maestros.

PREGUNTAS ORIENTADORAS

- ¿Su entidad tiene un protocolo estandarizado para la anonimización y protección de datos personales?
- ¿La entidad tiene establecidos protocolos para medir y evaluar la calidad de los datos de la entidad?
- ¿La entidad integra el proceso de gestión de datos con los procesos misionales e institucionales?
- ¿La entidad tiene un plan para mejorar la calidad de los datos asociados a indicadores clave de desempeño?
- ¿Los problemas o riesgos en gestión de datos son registrados en informes auditables y/o registros de riesgo?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
No hay claridad sobre el almacenamiento de datos estructurados y no estructurados en la entidad y/o no se sabe acceder a ellos.	Definir estrategias que rompan los silos de datos entre dependencias de la misma entidad. Se debe considerar la implementación de interfaces y consultas de datos cuando estos se encuentren en repositorios independientes.
No hay claridad sobre los protocolos de anonimización y protección de datos personales.	Documentar, conforme el marco de política y normativa definido en la planeación estratégica, todos los protocolos necesarios para anonimizar la información de carácter personal.

2.1.2.3. Capacidad operativa

Consiste en la materialización y ejecución de tareas definidas en la planeación táctica; ella permite transformar los recursos financieros, tecnológicos y humanos en productos y servicios para generar valor a partir de los datos. La capacidad operativa implementa acciones enmarcadas en los procesos definidos previamente en la gestión de los datos, y posibilita que en cada una de las etapas del ciclo de vida de los datos se pueda extraer el valor de estos. La ejecución de tareas de manera efectiva garantiza que se disponga de datos apropiados, de calidad y confiables para el desarrollo de proyectos e iniciativas de analítica de datos, que cumplan con el marco normativo determinado para la privacidad y seguridad de la información.

Entre las tareas operativas que integran la capacidad operativa se encuentran las siguientes: implementar mecanismos de captura y adquisición de datos externos; generar y adquirir datos y actualizar datos propios de la entidad; organizar los datos y determinar su disposición de almacenamiento; llevar a cabo la transformación y procesamiento de los datos para ser utilizados en procesos de analítica. Finalmente, las tareas operativas incluyen la implementación de técnicas de analítica de datos usando diferentes

lenguajes de programación como R, Python, SQL y Matlab y otros; la implementación y el despliegue de las herramientas de analítica de datos desarrolladas, así como la publicación y comunicación de los resultados obtenidos.

PREGUNTAS ORIENTADORAS

- ¿El mecanismo operativo para la generación de datos internos está definido, es estable y automatizado?
- ¿La entidad tiene procesos estandarizados para verificar la calidad de los datos que genera y captura?
- ¿Cómo están organizados los datos dentro de su entidad? ¿Existe un repositorio de información al que pueden acceder usuarios con determinados roles?
- ¿La información de la entidad está distribuida en silos o, por el contrario, la información se encuentra organizada y distribuida en repositorios?

RIESGOS	ESTRATEGIAS DE MITIGACIÓN
El uso de los datos en el proyecto genera dilemas éticos que no son gestionados en el ciclo de vida del proyecto.	Tener en cuenta los protocolos de consentimiento para el tratamiento de datos, anonimización de datos, la transparencia en el tratamiento de datos, la identificación de sesgos de discriminación o falta de representatividad de alguna población.
El uso de los datos puede generar riesgos de reidentificación de personas o sujetos.	Tener en cuenta las técnicas de anonimización de datos estructurados y no estructurados. (De-identificación, Anonimización de los datos, Encriptación de los datos, Aprendizaje de máquina con preservación de privacidad).
La calidad de los datos usados para el proyecto de analítica de datos no es suficiente para tomar decisiones a partir de ellas.	Se deben incorporar, desde la etapa de planeación táctica y estratégica, normativas y protocolos de calidad de datos que permitan establecer métricas para su evaluación y planes para su mejoramiento. A nivel operativo, los procesos de limpieza de los datos deben ser rigurosos, documentados y en la medida de los posible, automatizados y reproducibles.

2.1.3. RECOMENDACIONES GENERALES PARA FORTALECER LAS CAPACIDADES Y CONDICIONES HABILITANTES EN LAS ENTIDADES PÚBLICAS

Fortalecer las capacidades y condiciones habilitantes para la explotación de datos requiere de la puesta en marcha de acciones en la entidad, que en la medida de lo posible deben contar con el impulso de la alta dirección dado que involucran, como se mostró en la sección II.1.2, aspectos que superan la competencia exclusiva del área de tecnologías y sistemas de la información.

Por ello, avanzar en la definición y consolidación de una estrategia de explotación de datos implica cambios de paradigma en la entidad, principalmente los relacionados con la importancia de los datos como activo para la toma de decisiones. Este es el punto de partida para planificar e implementar acciones a fin de establecer una gobernanza de datos, en la que converjan los recursos físicos, humanos, tecnológicos y financieros hacia el cumplimiento de objetivos misionales institucionales y hacia la generación de valor público.

Para contar con el respaldo de la alta dirección, se recomienda vincular a los niveles directivos y hacedores de política de la entidad en la toma de decisiones basada en datos, para impulsar una estrategia de explotación de datos en aspectos decisivos de la entidad en cuanto a lo estratégico, táctico y operativo. También resulta fundamental visibilizar los beneficios de los proyectos que se hayan elaborado previamente en su entidad y/o en otras entidades públicas, esto permitirá que los directivos reconozcan el valor de los datos como activo para la toma de decisiones.

Los cambios organizacionales que requiera la entidad en materia de recursos humanos y tecnológicos y en planeación, se deben integrar a través de un cambio adaptativo para fortalecer las capacidades de la entidad de manera gradual. Es más importante avanzar paso a paso, planificando de manera cuidadosa las iniciativas de explotación de datos y tener un mecanismo de medición de los resultados de cada proceso.

La explotación y analítica de datos requiere priorizar la gestión del ciclo de vida de los datos en la entidad, y abordar la gobernanza de datos como un eje central de la gestión pública y administrativa. La gobernanza de datos permitirá, por una parte, definir todas las políticas, normativas y principios para la gestión del ciclo de vida de los datos desde la captura, almacenamiento, tratamiento, explotación y reutilización de datos.
Por otras parte, el aporte más importante de la gobernanza es que permite alinear las capacidades habilitantes para la explotación de datos con los objetivos estratégicos y misionales de la entidad.

En lo operativo es recomendable que la entidad establezca y documente las técnicas y procesos para salvaguardar la seguridad y la privacidad de los datos e implementar una gestión ética en la explotación de los datos. Ese conjunto de determinaciones permitirá fortalecer las iniciativas de analítica de datos, mientras que posibilita disminuir la incertidumbre y desconfianza para la compartición de datos. También se recomienda intercambiar experiencias y conocimientos entre científicos y analistas de datos de otras entidades públicas, de tal manera que se conviertan en un recurso compartido tano en el Gobierno nacional como en el territorial.

Por último, antes de trazar un plan u hoja de ruta para avanzar en el fortalecimiento de sus capacidades, conviene que la entidad identifique su nivel de madurez en cuanto a las capacidades y condiciones habilitantes que le permiten evolucionar en el desarrollo de una estrategia de analítica de datos de corto, mediano y largo plazo, siempre prestando atención a cada una de las capacidades habilitantes para avanzar de manera integral. Finalmente, al trazar una hoja de ruta concreta también han de definirse los responsables directos de las acciones en ella incorporadas.

CAPÍTULO 2.2

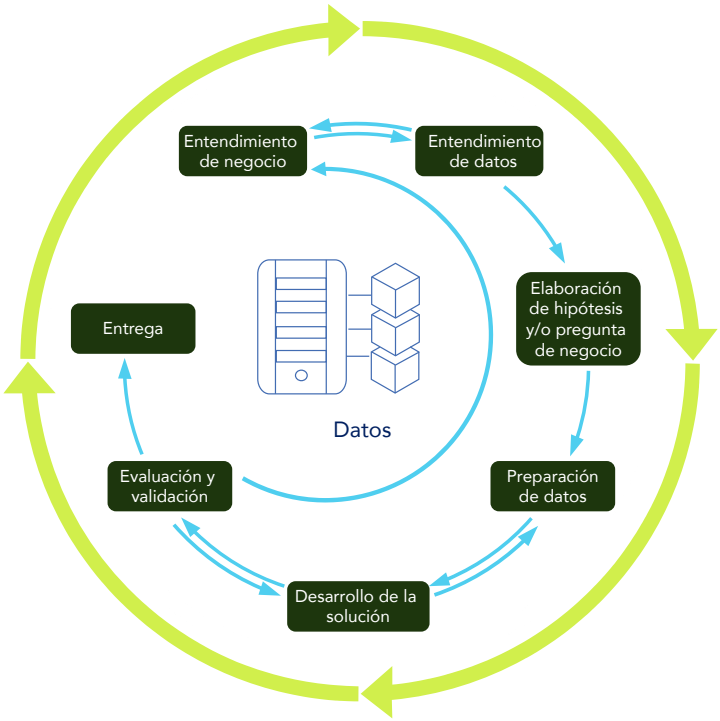
HOJA DE RUTA PARA EL DESARROLLO E IMPLEMENTACIÓN DE PROYECTOS DE ANALÍTICA DE DATOS

ESTE CAPÍTULO PRESENTA EL PROCESO DE DESARROLLO DE PROYECTOS DE ANALÍTICA DE DATOS, DESDE SU ENTENDIMIENTO HASTA SU VALIDACIÓN Y APROVECHAMIENTO. EL PROCESO COMPRENDE VARIAS ETAPAS QUE INVOLUCRAN A DIFERENTES ROLES DE LA ENTIDAD U ORGANIZACIÓN, LAS CUALES SE MUESTRAN EN UNA ADAPTACIÓN DEL MODELO CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING) (CHAPMAN, Y OTROS, 1999), METODOLOGÍA DISEÑADA PARA APLICARSE A PROYECTOS DE ANALÍTICA DE DATOS EN DIFERENTES INDUSTRIAS Y CONTEXTOS.

El proceso que se presenta a continuación incluye la modificación y adición de algunas etapas, considerando experiencias y lecciones aprendidas por la Unidad de Científicos de Datos (UCD) del Departamento Nacional de Planeación en el desarrollo de proyectos de analítica de datos.

A lo largo de este capítulo se describen cada una de las etapas y, en paralelo —en recuadros—, se presenta a manera de ejemplo un caso de uso que muestra cómo se llevó a cabo cada etapa en el proyecto *Detección y rastreo de inversión pública para el cumplimiento de los Objetivos de Desarrollo Sostenible (ODS)*, desarrollado por la UCD en el año 2019. Aunque las etapas que se muestran en la figura II.2-1 se inspiran en la metodología CRISP-DM, vale la pena anotar que esta no es la única en la industria, también existen alternativas como KDD (Knowledge Discovery in Databases) y SEMMA (Sample, Explore, Modify, Model, Assess), entre otras (Azevedo & Santos, 2008).

Figura II.2-1. Etapas para el desarrollo de un proyecto de analítica de datos. Adaptado de metodología CRISP-DM



Fuente: elaboración propia. Adaptación de Chapman, y otros, (1999).

Caso de uso:
Detección y rastreo de inversión pública para el cumplimiento de los ODS[8]

¿En estos recuadros se ilustran, por medio de un ejemplo aplicado, cada una de las etapas.

Conozca más sobre este y otros proyectos llevados a cabo por la UCD ingresando a:

8. *ODS:* Objetivos de Desarrollo Sostenible.

2.2.1. ENTENDIMIENTO DE NEGOCIO

Lo primero que debe hacerse en cualquier proyecto, sea de analítica de datos o no, es identificar y entender exactamente qué es lo que se desea lograr. Se debe tener una compresión clara de cuál es el problema por resolver mediante analítica de datos, cuáles son las metas, las restricciones, el alcance y el impacto de la solución. La falta de claridad en estos puntos puede causar pérdida de tiempo desarrollando soluciones que al final no son de interés, ni resuelven el problema. Por lo tanto, es fundamental entender las necesidades y problemáticas del negocio con el apoyo de personas que tengan un dominio sobre el tema.
Lo trascendental de tener estas precisiones es que pueden garantizar, en gran medida, el éxito del proyecto.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

ENTENDIMIENTO DEL NEGOCIO: INVERSIÓN PÚBLICA EN COLOMBIA

El caso de uso se enmarcó en el sector de inversión pública.
El acompañamiento de expertos temáticos del sector fue clave para entender que los proyectos tienen diferentes fases de ejecución —programación, ejecución, operación, seguimiento y evaluación posterior— y se financian con diversas fuentes de recursos —Presupuesto General de la Nación (PGN), Sistema General de Regalías (SGR), Sistema General de Participaciones (SGP), Cooperación Internacional y recursos de las entidades territoriales, entre otros—.

Por ello, el proyecto se desarrolló con el acompañamiento del equipo de la Secretaría Técnica de la Comisión ODS del DNP y por el equipo que diseñó y administra la base de datos de Mapa Inversiones, principal insumo de datos para el proceso realizado.

2.2.2. ENTENDIMIENTO DE LOS DATOS

Después de entender el negocio, la segunda etapa consiste en el entendimiento de los datos. Entender su naturaleza, es decir, conocer las fuentes de información, si los datos son públicos, privados o con restricciones; saber la periodicidad con la que se actualizan; el tipo de información que se reporta, si son datos estructurados y no estructurados; y verificar si la es información apropiada para cumplir las metas e impactos establecidos en la primera etapa. Con base en este entendimiento es posible, incluso, llegar a descubrir fallas en el entendimiento del negocio, lo que ayuda a replantear los objetivos y los planes desde el inicio sin desaprovechar esfuerzos.

Cabe anotar que en esta etapa se consolidan las fuentes de información, insumo esencial para el desarrollo de los proyectos de analítica de datos. En este punto aún no se efectúa ningún tipo de proceso o transformación sobre los datos.

Una vez se tenga esta consolidación, estos son algunos puntos para atender:

Identificar la unidad de análisis

Qué representa cada registro en la base de datos. Por ejemplo, en bases de datos poblacionales un registro puede representar una persona, un hogar, un municipio, entre otros; en texto, puede ser una palabra, un documento o alguna medida de similitud.

Comprender la metodología que se usó para recolectar los datos

Al hacer preguntas como ¿qué método de muestreo se utilizó?, ¿la información está delimitada en un tiempo determinado?, ¿los datos corresponden a un área geográfica? y otras preguntas del mismo tipo.

Identificar la herramienta de recolección de datos

Si esta fue hecha mediante un formulario, encuesta, un algoritmo, información obtenida por sensores, entre una gran variedad de métodos.

Los puntos anteriores pueden dar indicios de posibles errores debidos a anomalías en los datos como asuntos faltantes, errores de formato, datos duplicados y datos atípicos, entre otras condiciones.

Tales anomalías pueden representar un error o ser normales en el conjunto de datos; lo anterior cambia de acuerdo con cada conjunto de datos particular, por ello es necesario tener un buen entendimiento de los datos y del contexto para determinar si hay errores de información.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

ENTENDIMIENTO DE LOS DATOS: SUIFP Y APC

Para este proyecto se estuvieron disponibles dos bases de datos: la información de Mapa Inversiones y la de la Agencia Presidencial de Cooperación (APC). En ambas fuentes, cada registro tiene información asociada a un proyecto de inversión; por ende, se definió que cada proyecto de inversión sería una unidad de análisis.

Se contó con el censo de los proyectos públicos financiados con recursos de SGR, PGN y Cooperación Internacional, dado que todo proyecto financiado con esas fuentes debía estar registrado en alguna de las bases de datos.

Se tomaron los proyectos de inversión en una ventana de tiempo desde 2012 hasta 2019, incluidas vigencias futuras hasta 2026, aunque en el caso de los recursos de cooperación internacional solo se contaba con información hasta 2018. Se identificaron las variables de interés para cumplir con las metas fijadas en el entendimiento del negocio: BPIN (ID del proyecto), fuente de recursos, año, valor programado, sector, municipio de ejecución de los recursos y las columnas de datos textuales —título, objetivos, descripción, etc.— que contenían información de cada proyecto. También se identificó que los campos de texto disponibles permitían identificar, para la mayoría de los proyectos, si estos se relacionaban o no con uno o más ODS.

2.2.3. FORMULACIÓN DE LA HIPÓTESIS O PREGUNTA DE NEGOCIO

Tras finalizar las etapas de entendimiento del negocio y entendimiento de los datos, surge una de las etapas de mayor dificultad y estrechamente relacionada con las necesidades y problemáticas identificadas en el entendimiento de negocio: plantear una pregunta específica a la que se vaya a dar respuesta con el desarrollo del proyecto. Esa pregunta debe contar con tres características (Mattick, Johnston, & Croix, 2018):

1 **Relevancia.**
Evalúe si es importante para la entidad responder a esta pregunta, sea por temas estratégicos, táctico u operativos.

2 **Originalidad.**
Revise que al responder la pregunta se brinde nueva información o nuevos insumos y que no se esté duplicando trabajo realizado previamente.

3 **Rigor.**
Verifique que la pregunta pueda responderse con ayuda de los datos y técnicas disponibles.

El proceso de formulación puede facilitarse mediante el planteamiento de una pregunta inicial que se acota a partir de aspectos geográficos, temporales, sectoriales, de contexto y de enfoque. También se precisan las preguntas e hipótesis planteadas que deben ser ampliamente discutidas, socializadas y aceptadas por los expertos temáticos antes de ser tomadas como el pilar principal del proyecto de analítica de datos.

Un escenario alternativo que igualmente puede presentarse en esta etapa es que se tienen unos activos de información disponibles, que se perciben valiosos o se quieren aprovechar, pero no hay una necesidad o pregunta identificada. En tal caso, pueden plantearse objetivos de tipo descriptivo y exploratorio para ganar un mejor entendimiento de los datos disponibles. Otra opción es definir o identificar necesidades y objetivos que puedan alcanzarse a partir de la información existente. A continuación se presentan preguntas que pueden ser de utilidad para dar un manejo acertado a la situación descrita:

- *¿El contenido y la calidad de los datos disponibles es adecuado para plantear un proyecto de analítica o explotación de datos?*
- *¿Qué necesidades se pueden atender con los datos disponibles?*
- *¿Los problemas o necesidades encontradas son realmente significativos para resolver en la entidad?*
- *¿Es relevante, o representa beneficios potenciales para la entidad, dar respuesta a esa problemática o necesidad?*
- *¿Es un proyecto de analítica de datos la mejor alternativa para responder o atender las necesidades identificadas?*

Detección y rastreo de inversión pública para el cumplimiento de los ODS

FORMULACIÓN DE LA PREGUNTA: ¿QUÉ SE DESEA RESPONDER U OBTENER CON ESTE PROYECTO?

Se parte de una necesidad identificada previamente por el grupo de ODS del DNP. Las metas asociadas a cada ODS consideran unos indicadores ya constituidos para hacer seguimiento y verificar el avance y cumplimiento de cada uno de ellos. Sin embargo, es necesario contar con indicadores financieros complementarios que permitan tener una visión global sobre cómo la inversión de recursos públicos se orienta hacia la consecución de cada ODS. Con base en la problemática planteada y en los datos disponibles, se definió la siguiente pregunta:

¿Cuál es el monto total de recursos que acumulan los proyectos de inversión que se encuentran en alineación con cada uno de los ODS?

A partir de del interrogante planteado, se definieron para el proyecto los siguientes objetivos:

Objetivo general

Analizar la inversión pública realizada en alineación con los ODS a través de la aplicación de analítica de datos.

Objetivos específicos

1. Consolidar datos sobre los proyectos de inversión y los ODS, mediante la realización de consultas en bases de datos públicas y del DNP con información estructurada y no estructurada.

2. Etiquetar los proyectos de inversión como relacionados o no relacionados con cada uno de los ODS a través de la aplicación de técnicas de análisis de texto y procesamiento de lenguaje natural.

3. Presentar los resultados de manera organizada, con enfoque territorial y enfoque sectorial a través del desarrollo de una herramienta de visualización.

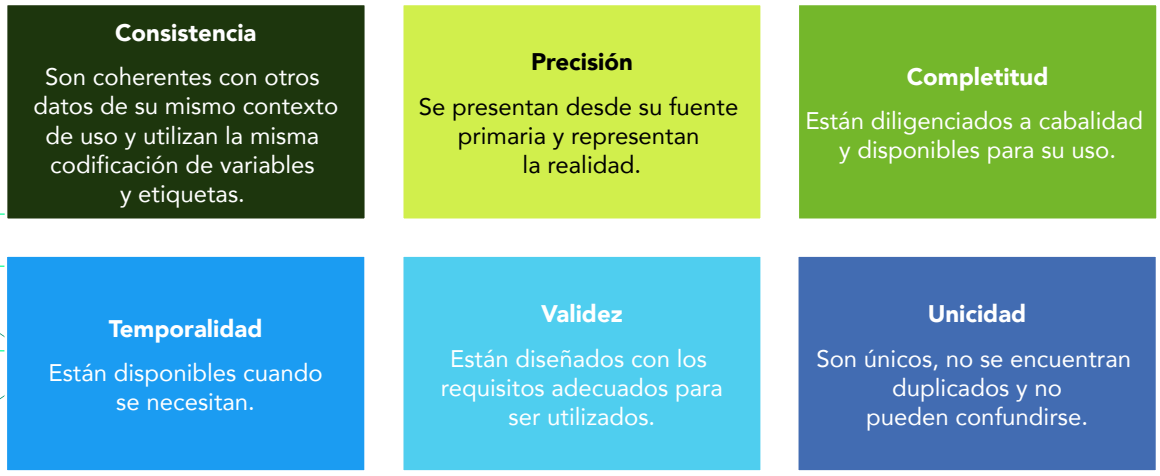
2.2.4. PREPARACIÓN DE LOS DATOS

Es la etapa en la que los roles de científicos de datos e ingenieros de datos pasan la mayor parte de su tiempo. Una encuesta de la revista *Forbes* en 2016 afirma que esta etapa toma en promedio el 80% del tiempo de los científicos de datos en sus análisis (Press, 2016). Después de haber seleccionado las fuentes de datos en las etapas anteriores, estos deben adecuarse o procesarse de forma que sean aptos para las etapas de exploración, modelamiento y análisis. Las adecuaciones pueden ir desde procesos sencillos como una normalización de los datos hasta tareas más complejas como transformaciones en ellos, tratamiento de valores faltantes, tratamiento de valores atípicos, entre muchas otras. Parte de las tareas más usuales para la preparación de los datos se describen a continuación —cabe aclarar que la aplicación de una o varias de estas acciones depende de cada proyecto—.

2.2.4.1. Revisión de calidad y valor de los datos

La calidad de los datos es crucial, pues ella permite medir la confiabilidad de la información y qué tanto es capaz de cumplir con los propósitos en un contexto en particular. Hay varias características que habilitan medir la calidad de un conjunto datos; entre las más usadas por la comunidad de ciencia de datos están las siguientes: precisión, temporalidad, completitud, unicidad, consistencia y validez. En la figura II.2 -2 se describen de forma breve cada una de ellas.

Figura II.2-2. Características para medir la calidad de los datos



Fuente: adaptado de (Lean-Data, 2018).

2.2.4.2. Limpieza de datos

Comprende el proceso de detectar y corregir —o eliminar— registros corruptos o inexactos en las bases de datos. El proceso se puede definir en cuatro pasos, a saber (Abdallah, Du, & Webb, 2017): 1) definición e identificación de errores; 2) corrección de los errores por reemplazo, modificación o eliminación; 3) documentación de los tipos de error y su ubicación; 4) medición y verificación de que los datos, luego de la limpieza efectuada, cumplan los requisitos para ser utilizados como insumo, y así alcanzar las metas del proyecto de analítica de datos.

2.2.4.3. Enriquecimiento e integración de datos

Consiste en identificar fuentes de información adicionales para potenciar o complementar los datos disponibles actualmente y, de esta forma, mejorar el alcance del proyecto. El proceso de integración es la combinación de las fuentes de información adicionales con las bases de datos principales. La integración puede imponer el procesamiento adicional de una o varias fuentes de información para asegurar que sean compatibles entre sí.

2.2.4.4. Adecuación de variables

La adecuación de las variables —también conocidas como *columnas o características*— de una base de datos puede traer muchos beneficios para el desempeño final del proceso de modelamiento y análisis. Unos de los beneficios son estos: mejor interpretación de los datos; disminuir varianzas dentro de cada variable; mitigar efectos de datos corruptos o extremos; cambiar la distribución de las variables, y mejorar la compatibilidad del conjunto de datos con supuestos establecidos en la formulación de la hipótesis. Tales adecuaciones incluyen la normalización de los datos (mínimo–máximo o estandarización normal), transformaciones lineales o no lineales, binarización y discretización, entre otras.

2.2.4.5. Creación de características

Esta etapa corresponde al cálculo o la derivación de nuevas variables a partir de las variables originales. Este proceso se diferencia de la adecuación de variables en cuanto a que no busca simplemente cambiar su representación, sino extraer información de ellas para crear una nueva. Los procesos de extracción de características pueden variar, desde la extracción de información específica de una variable —como la hora o el día del mes de una variable de fecha— hasta la creación de una variable a partir de dos o más variables existentes —como calcular el beneficio a partir de variables de ingresos y costos—. El beneficio principal de este paso es la mejora en los modelos de estimación y predicción.

2.2.4.6. Reducción de dimensionalidad

Consiste en reducir la dimensión de la base de datos, o el número de variables, a un número menor. Las razones para emplear esta técnica pueden variar desde una mejora en la visualización de una base de datos, mejorar procesos de agrupamiento, agilizar los tiempos de entrenamiento y estimación, eliminar información redundante o de ruido de la base de datos o evitar sobreajuste en modelos de aprendizaje de máquina. Además, la reducción de dimensionalidad también se utiliza en el proceso de selección de características como una manera automatizada de escoger variables para incluir en un modelo. Entre técnicas de reducción de dimensionalidad se pueden listar las siguientes: PCA, análisis de factores, LDA, Isomap, MDS y t-SNE. La escogencia de la técnica de reducción de dimensionalidad depende de la naturaleza de los datos, y seleccionar una u otra puede mejorar los resultados finales.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

PREPARACIÓN DE LOS DATOS: RECOLECCIÓN Y CRUCE DE INFORMACIÓN, Y PROCESAMIENTO DE DATOS TEXTUALES Y FINANCIEROS.

Enriquecimiento de datos. En este proyecto fue necesario consolidar información de DNP junto con información no estructurada de fuentes externas.

Integración y selección de variable. Para cada proyecto de inversión, se consultó sobre tres tablas de SQL asociadas a la base de datos de Mapa Inversiones, y se obtuvieron un total de 14 variables para cada uno.

En estas consultas se delimitó la información a los registros provenientes de SUIFP PGN y SUIFP SGR, que corresponden al presupuesto aprobado —y no al presupuesto ejecutado— para los proyectos de inversión, sean estos financiados con recursos de PGN, de SGR, de SGP, de las entidades territoriales, de empresas públicas o con recursos privados.

Preparación de los datos. Obtenidos los insumos de SUIFP para los proyectos de inversión, se consolidó la información de la APC para los proyectos de cooperación internacional. En este caso, se tomó la base con corte a 2018 que proveyó por el grupo ODS del DNP; se procedió a totalizar los aportes registrados en la base de datos, independientemente del actor que efectuó el aporte, y se extrajo un valor total destinado a cada proyecto de cooperación internacional.

Finalmente, se utilizaron y procesaron textos sobre ODS sobre las metas y objetivos definidos para Colombia en la Agenda 2030. Los textos se complementaron con las cartillas de "Por qué es importante" de cada ODS, así como con su correspondiente introducción y sus datos destacables, todos ellos disponibles en la página web de las Naciones Unidas.

2.2.5. DESARROLLO DE LA SOLUCIÓN

En esta etapa se comienza a dar respuesta a las necesidades planteadas en el entendimiento del negocio y el entendimiento de los datos. Aquí se debe elegir la técnica de análisis o modelamiento que mejor se ajuste al problema o al objetivo que se desea alcanzar. Esto puede variar desde soluciones que impliquen dar un reporte, visualizaciones de los datos y tableros de control hasta técnicas más complejas de análisis como el desarrollo de modelos estadísticos, de minería de datos y/o aprendizaje automático. La solución podrá variar mucho según las necesidades puntuales de cada proyecto en complejidad tanto *analítica* —yendo desde estudios descriptivos y exploratorios hasta modelos predictivos y prescriptivos— como *tecnológica* —sea por los requerimientos de consumo de datos o por el despliegue de herramientas—.

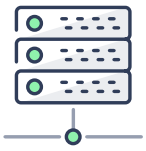
Conviene tener presente que la solución por desarrollar está estrechamente relacionada con el tipo y la naturaleza de los datos. La calidad, la cantidad y la granularidad de los datos disponibles determinará, en gran medida, cómo estos pueden ser aprovechados y qué tipo de técnicas y modelos pueden aplicarse. De acuerdo con el tipo de datos —estructurados, semiestructurados, no estructurados, secuenciales, geográficos, etc.— también habrá unas técnicas de procesamiento, análisis y modelamiento más apropiadas que otras. Finalmente, en cuanto al desarrollo de modelos, la disponibilidad o no de datos etiquetados —registros acompañados de una columna adicional que contiene el valor correspondiente de una variable objetivo o de interés— afectará también el tipo de modelos que se pueden entrenar y ajustar.

2.2.5.1. Desarrollo y entrenamiento de modelos

Esta etapa es necesaria cuando se ha identificado que la solución del problema se llevará a cabo mediante técnicas de aprendizaje automático. Lo primero en este caso es definir de forma clara qué tipo de solución será. Por ejemplo:

Problemas de clasificación como tipificación de imágenes, detección de fraude o detección de spam.

Problemas de regresión como estimación de crecimiento de una población, predicciones meteorológicas o del precio de una acción.



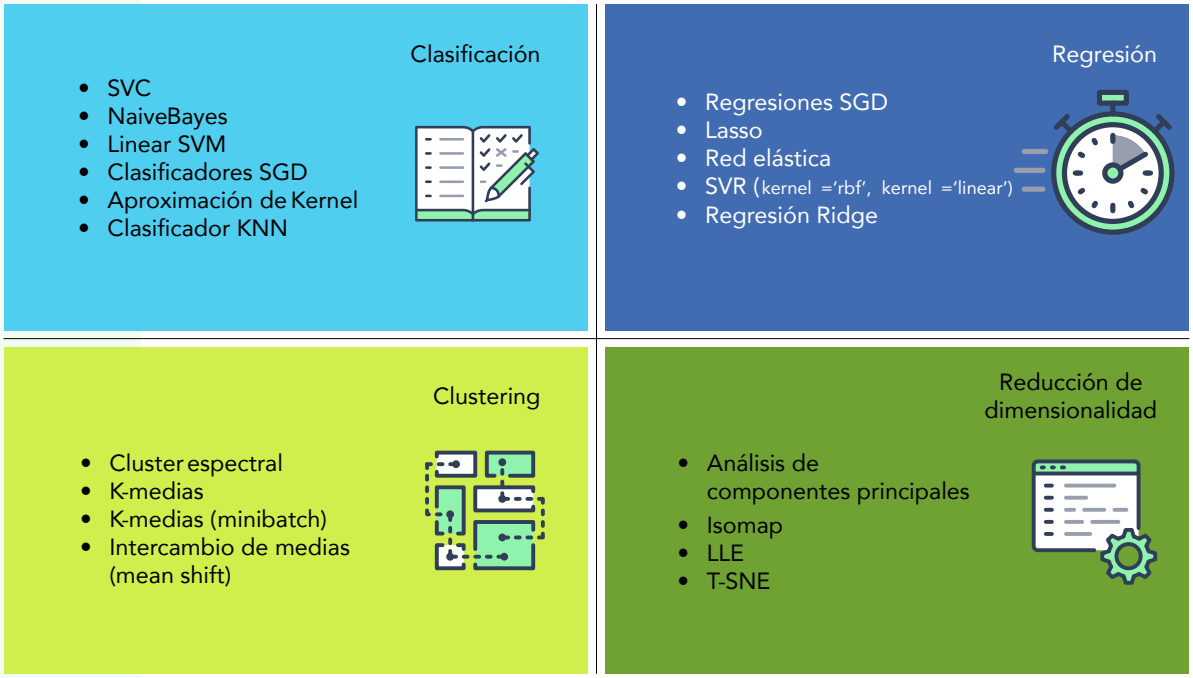
Problemas de agrupamiento o segmentación de datos como el *clustering* de documentos por temas similares o la segmentación de usuarios.

Problemas de aprendizaje por refuerzo como entrenar un autómata para moverse en un ambiente no del todo conocido o controlado.

Los problemas de clasificación y regresión se pueden resolver con modelos de aprendizaje supervisado. En esos modelos se tiene un conjunto de variables predictoras, con las que se pretende aproximar una o más variables objetivo con el menor error posible. Este tipo de modelos son los más utilizados por diferentes industrias en la actualidad. Los dos últimos problemas listados anteriormente pueden resolverse mediante modelos de aprendizaje no supervisado y aprendizaje por refuerzo, en su orden.

Una de las librerías más populares para el desarrollo y entrenamiento de modelos en Python es *scikit-learn*. En la figura II.2-3, dependiendo de la naturaleza de la tarea, se muestra una hoja de ruta para la selección de modelos de aprendizaje de máquina propuesto por los autores de *Scikit-learn* (Pedregosa, y otros, 2011).

Figura II.2-3. Algoritmos de aprendizaje automático recomendados de acuerdo con la naturaleza de la tarea



Fuente: adaptado de Choose the best estimator (Sklearn, 2016).

2.2.5.2. Sintonización y validación de modelos supervisados

Cuando se entrena un modelo de aprendizaje supervisado, generalmente hay una etapa de sintonización de los hiperparámetros del modelo, en la que se prueban varios valores configurables para uno o varios de ellos, buscando la combinación que produzca el mejor desempeño. Por ejemplo, para un clasificador KNN (K-Nearest Neighbors), se debe elegir el número óptimo de vecinos por considerar y la distancia de similitud por utilizar, para obtener mejores desempeños de clasificación.

Para este proceso se aconseja dividir los datos, después de etapa de limpieza y adecuación, en tres conjuntos: *entrenamiento, validación y prueba*. El propósito de esta división es entrenar el modelo sobre el conjunto de

entrenamiento y evaluar su desempeño en el conjunto de validación, para detectar y evitar sobreajustarlo. Finalmente, una vez se ha entrenado el modelo con la mejor combinación de parámetros, se utiliza el conjunto de prueba para evaluar la capacidad que tiene el modelo para generalizar ante datos que no ha visto antes.

El porcentaje de datos asignado a cada conjunto depende de la cantidad de registros disponibles en la base de datos. Aunque la distribución de los grupos de entrenamiento, validación y prueba pueden variar de acuerdo con el tipo de modelo y el problema específico, los siguientes lineamientos dan una idea general de cómo distribuir los datos disponibles, según su cantidad:

1

Menos de 100 registros

Es una cantidad muy baja de datos. Es poco probable que un modelo entrenado con estos datos tenga un buen desempeño en problemas con un mínimo grado de complejidad.

2

De 100 a 5.000 registros

La cantidad de datos disponibles para entrenamiento sigue siendo baja, por lo que se recomienda destinar la mayor cantidad de registros —un 90%— al entrenamiento de los modelos, y dejar el 10% restante para validación. En estos casos no habría un conjunto de prueba.

3

De 5.000 a 20.000 registros

Se recomienda una distribución de entrenamiento al 80%, una validación al 10% y una prueba al 10%.

4

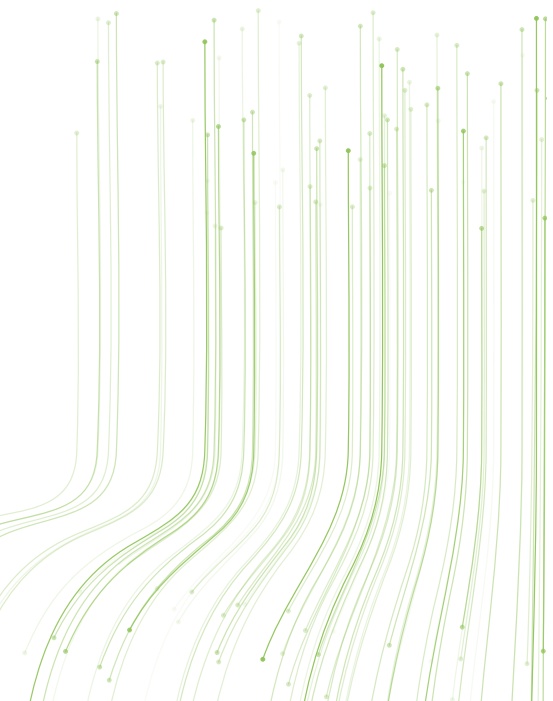
De 20.000 a 100.000 registros

Se recomienda una distribución de entrenamiento al 70%, una validación al 15% y prueba al 15%.

5

Más de 100.000 registros

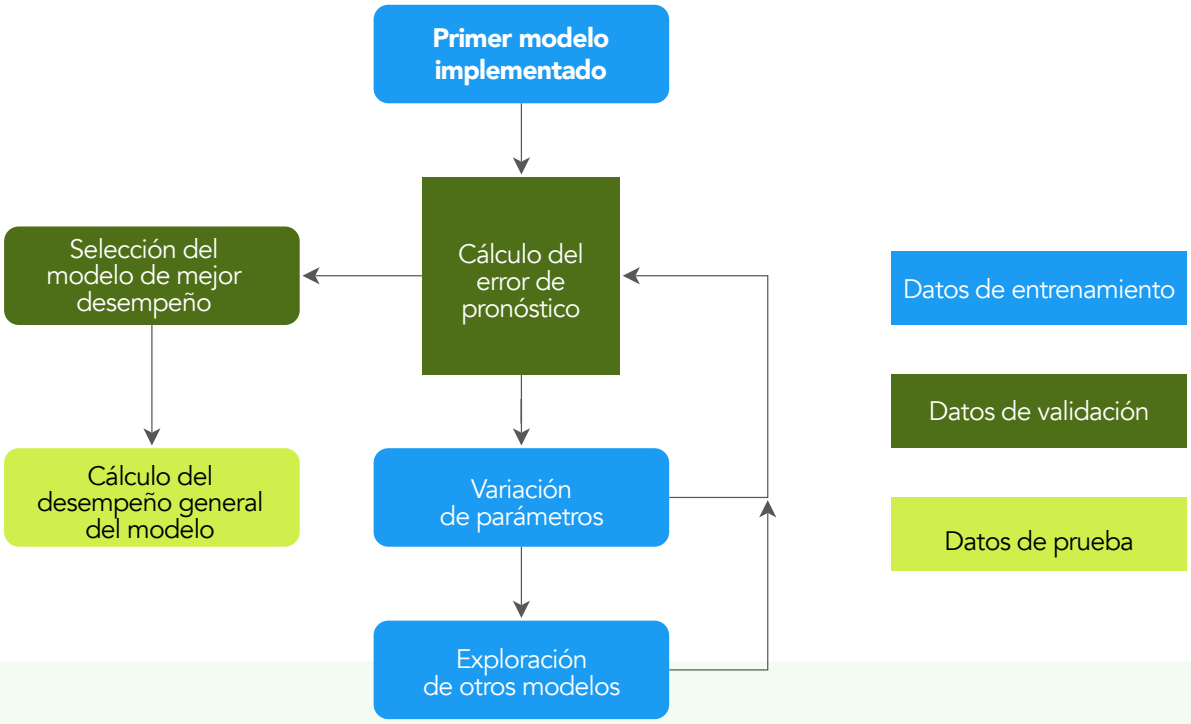
A medida que se tiene un conjunto más grande de datos, se recomienda destinar la mayoría de los registros al entrenamiento de los modelos, para poder tener modelos más robustos y de mejor desempeño. En lo posible, se recomienda tener grupos de validación y prueba de 20.000 registros cada uno, y dejar el resto de los registros disponibles para entrenamiento.



Por lo general, esta etapa inicia entrenando un modelo sencillo para obtener un desempeño base sobre el conjunto de entrenamiento. Luego, en un proceso iterativo, se van sintonizando hiperparámetros y probando otros modelos hasta encontrar el de mejor desempeño. Este proceso se ilustra en la figura II.2-4.

Más adelante, en la sección II.2.6, se describen las métricas que se utilizan para medir los desempeños de los modelos, y cómo estas ayudan a la selección de la opción más apropiada.

Figura II.2-4. Diagrama de flujo para desarrollar modelos de aprendizaje automático



Fuente: elaboración propia.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

DESARROLLO DE LA SOLUCIÓN (MODELAMIENTO): CÁLCULO DE SIMILITUDES ENTRE TEXTOS Y DETERMINACIÓN DE UMBRALES DE CLASIFICACIÓN

Inicialmente, para el desarrollo de la solución para el proyecto "rastreo de recursos para el cumplimiento de los ODS" se propuso efectuar un etiquetado de algunos proyectos a partir de la identificación de palabras clave, y utilizar los proyectos etiquetados para entrenar un modelo de aprendizaje de máquina que predijera, a partir del texto que describe cada proyecto, la probabilidad de que se encontrara alineado con cada ODS. Sin embargo, los resultados preliminares con dicha metodología no fueron satisfactorios, lo que llevó a considerar enfoques diferentes.

Este es un caso común en los proyectos de analítica, por lo que se reitera la importancia de tener una retroalimentación constante de la metodología y de los resultados preliminares, tanto con personas del sector como con personas que conozcan las técnicas de analítica de datos, pues tal proceso facilita orientar mejor la metodología y evita incurrir en esfuerzos grandes que pueden ser desgastantes y que pueden conducir a no obtener los resultados esperados.

Así, la metodología con la que se desarrolló la versión final de este proyecto consistió en el cálculo de una medida de similitud entre textos de ODS y textos de proyectos de inversión, seguido del ajuste de modelos de mixturas para etiquetar los proyectos como alineados o no con cada uno de los ODS. Esta metodología no solo fue más sencilla, sino que mostró generar resultados mucho más satisfactorios y acordes con lo esperado.

2.2.5.2. Desarrollo de herramientas

Por lo común, el desarrollo de un modelo o análisis a partir de los datos no es suficiente para que este pueda ser debidamente aprovechado por los usuarios finales. Por ellos, es necesario llevar a cabo algún tipo de despliegue del trabajo efectuado. La etapa de desarrollo de herramientas es donde se diseña y desarrolla lo que se desplegará, si es necesario. En este punto se define y desarrolla el modo como se presentarán los resultados del proyecto al usuario, ya sea a través de un tablero de visualización, una aplicación, un proceso *batch*, un reporte, una tabla o un dato puntual, entre otras opciones. Algunos *frameworks* (tecnologías y entornos de trabajo) que pueden ser útiles para el desarrollo de este tipo de herramientas se encuentran en la figura II.2-5.

Figura II.2-5. Frameworks útiles para el desarrollo de herramientas de analítica

Fuente: elaboración propia.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

DESARROLLO DE LA SOLUCIÓN (HERRAMIENTAS): DESARROLLO DE APLICACIÓN WEB EN SHINY PARA VISUALIZAR LOS RESULTADOS

Se desarrolló un aplicativo en el *framework* de Shiny (R) para visualizar los resultados. En el aplicativo, el usuario escoge el ODS que desea analizar y puede modificar el umbral de clasificación, lo que permite mayor flexibilidad si se desean aumentar o disminuir las tasas de falsos negativos y falsos positivos. Definido el umbral, el usuario puede filtrar los proyectos de inversión por fuente de recursos, sector —para el caso de "SGR" y "Territorial"— y vigencia.

La aplicación muestra al usuario una tabla con los proyectos organizados de mayor a menor similitud con respecto al ODS elegido. También muestra el valor aprobado para el proyecto en cada una de las vigencias escogidas. De manera más agregada, la aplicación reporta, según la información disponible, el monto total de los proyectos clasificados —con los filtros y el umbral escogidos—, un mapa de calor con los montos de inversión por departamento —enfoque territorial—, una serie de tiempo con los recursos por vigencia, un gráfico de barras con los recursos por sector —enfoque sectorial— y un gráfico de torta que muestra la comparación de la inversión programada en alineación con el ODS contra el total de recursos aprobados.

2.2.6. EVALUACIÓN Y VALIDACIÓN

2.2.6.1. Evaluación y validación de modelos

En esta sección se presentan algunas de las técnicas y métricas comúnmente usadas para la evaluación y validación de los modelos, teniendo en cuenta el tipo de aprendizaje —supervisado, semisupervisado, o no supervisado— y el tipo de tarea —regresión, clasificación, agrupamiento, otros—.

Como se mencionó en la sección II.2.5.1, al validar un modelo de aprendizaje de máquina se recomienda dividir los datos en tres conjuntos: entrenamiento, validación y prueba, acción sujeta a la cantidad de datos disponibles. Estos subconjuntos de datos interactúan en la selección del mejor estimador mediante un proceso iterativo, como se muestra en la figura II.2-4, en donde en cada iteración se evalúa una métrica de error de la predicción. A continuación, se describen las métricas que pueden utilizarse considerando la naturaleza de la tarea y el tipo de modelo.

2.2.6.2. Modelos supervisados

Este tipo de modelos se utilizan cuando se tienen datos asociados a una etiqueta de clase o valor; por ejemplo, en problemas de detección de objetos las imágenes cuentan con unas etiquetas que indican los objetos que están presentes en cada imagen y sus respectivas coordenadas. Otro ejemplo es la detección de correos que contienen *spam*; en este caso, para cada correo del conjunto de entrenamiento se tendrá una variable binaria que indicará si dicho correo es de spam o no. Para este tipo de modelos existen dos tareas principales:

Regresión

Las métricas más populares son RSME (*Root Mean Square Error*), MAE (*Mean Absolute Error*) y *R square*. En Swalin (2018) se puede encontrar una guía práctica del uso de estas métricas para problemas de regresión.

Clasificación

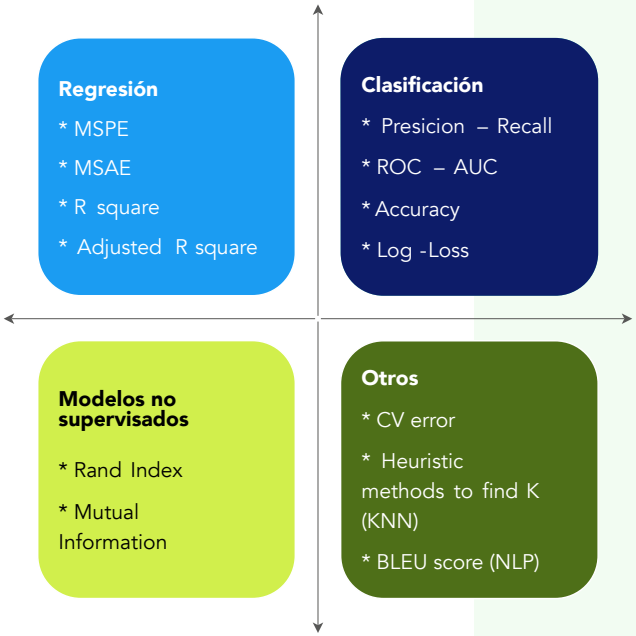
Las métricas que se utilizan usualmente son *Accuracy, Precision-Recall*, curvas ROC-AUC, *Logarithmic loss y F1 score*. Ejemplos de selección de la métrica adecuada dependiendo el problema se pueden encontrar en Swalin (2018) y Mishra (2018).

Un problema adicional que puede surgir en los modelos de clasificación es el desbalance entre las categorías de entrenamiento; en otras palabras, que en una o en varias categorías de la base de datos haya un número de registros mucho menor al que hay para otras de ellas. Lo anterior agrega más condiciones para escoger la métrica correcta, pues medidas como el *Accuracy* pueden hacer caer en una mala interpretación del desempeño del estimador. En Brownlee (2020), se describe cómo analizar e interpretar las métricas para este tipo de problemas.

2.2.6.3. Modelos no supervisados

El aprendizaje no supervisado es un tipo de aprendizaje automático que busca descubrir patrones no detectados previamente en un conjunto de datos en el que no hay datos etiquetados, a diferencia del aprendizaje supervisado que generalmente utiliza datos etiquetados por humanos. Las métricas para este tipo de aprendizaje se basan en correlaciones entre los datos y teoría de la información. Las métricas más utilizadas son el *Rand Index* y *Mutual Information*, pero existe una gran variedad de métricas según de la naturaleza de los datos. En Palacio-Niño & Berzal (2019), se puede encontrar un resumen y una descripción de las métricas más utilizadas en tareas de aprendizaje no supervisados. En la figura II.2-6 se muestra un resumen de las métricas que más se usan en la evaluación y validación de modelos.

Figura II.2-6. Métricas de desempeño para modelos de machine learning.



Fuente: adaptación de Swalin, (2018).

Como ya se mencionó, al igual que en los conjuntos de datos con desbalance de clases, la interpretación de cada métrica puede dar pistas de aciertos o errores en el proceso de entrenamiento de un modelo; por ello, en muchas ocasiones se recomienda no interpretar las métricas como un valor numérico.

En Amazon (2020) hacen una descripción de cómo se debe interpretar cada una de las métricas para modelos de *machine learning*.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

EVALUACIÓN Y VALIDACIÓN (MODELAMIENTO):
CLASIFICACIÓN DE PROYECTOS POR ODS

En la primera metodología con que se abordó el problema utilizó las medidas de precisión y *recall* para escoger los mejores modelos de clasificación. La mayoría de los modelos no superaron el 60 % de precisión en la clasificación, por lo que se pudo intuir que los resultados no serían satisfactorios, algo que se confirmó con la caracterización de los proyectos clasificados con esta metodología.

Para el caso de la segunda metodología, como no se realizó un etiquetado de categorías, no se pudieron utilizar métricas para evaluar modelos supervisados. Por ello, se evaluó el resultado con métricas de separación de clases como *mutual information* y el estadístico K-S (Kolmogórov-Smirnov), seguido de una validación con los expertos.

2.2.6.4. Evaluación y validación de herramientas

En la sección II.2.5.2, se mencionó que los proyectos y los productos de analítica de datos en muchas ocasiones se hacen disponibles a través de herramientas, tableros de visualización o aplicativos que serán utilizados directamente por los usuarios finales. Por tal motivo, es importante practicar pruebas y validaciones de cada herramienta o producto antes de su puesta en marcha y uso directo. Estas pruebas tienen como objetivo hacer tres verificaciones, a saber:

1

El correcto desempeño de los módulos

Que no existan errores de funcionamiento en rutinas y subrutinas de la herramienta.

2

Funcionalidades que componen la herramienta

Comprobar que la herramienta hace lo que debe hacer.

3

Experiencia del usuario

Verificar que la interfaz o interacción con la herramienta sea amigable e intuitiva, de forma que los usuarios puedan utilizarla de forma correcta.

Estas validaciones pueden agruparse así:

1

Tipos de pruebas por su ejecución

Ejecución manual o automática.

2

Por el enfoque de la validación

Pruebas de caja negra —se validan entradas y salidas—, pruebas de caja blanca —en cada proceso del funcionamiento se conocen las variables— y pruebas aleatorias —se validan al azar los componentes de la herramienta—.

3

Clasificación de las pruebas según lo que verifican

Se dividen en funcionales y no funcionales. *Las pruebas funcionales* se basan en la ejecución de la herramienta; hay distintos tipos de pruebas funcionales, como las pruebas Alpha, pruebas Beta y pruebas de aceptación. *Las pruebas no funcionales* tienen como objetivo verificar un requisito que especifica criterios que pueden usarse para juzgar la operación de la herramienta; por ejemplo, disponibilidad, accesibilidad, usabilidad, mantenibilidad, seguridad, rendimiento y otras. Las pruebas más recurrentes de este tipo son las pruebas de seguridad, las pruebas de estrés y las pruebas de usabilidad.

Algunas herramientas populares de código abierto para evaluación y validación de productos de software son:

Para prueba funcionales

Selenium, Soapui, Watir.

Para pruebas de carga y rendimiento

Jmeter, Gatting y FunkLoad.

Sin embargo, la elección de las herramientas para validar el correcto comportamiento y desempeño de un producto en específico dependerá de varios aspectos como las tecnologías utilizadas para desarrollarlo, su complejidad y el alcance del producto desarrollado, entre otros.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

EVALUACIÓN Y VALIDACIÓN DE LA SOLUCIÓN (HERRAMIENTAS): PRUEBAS DEL APLICATIVO WEB

El aplicativo desarrollado fue sometido a pruebas de uso por parte de los miembros del equipo de la UCD y de usuarios externos. Ese proceso permitió, además de validar la correcta funcionalidad del aplicativo, retroalimentar la experiencia del usuario al utilizarlo. Dado que no se esperaba que el aplicativo contara con muchos usuarios simultáneos, no se realizaron pruebas de carga, capacidad y estrés, las cuales son muy frecuentes para el despliegue de aplicativos web.

2.2.7. ENTREGA

Es el paso final en el desarrollo de un proyecto de ciencia de datos. De acuerdo con el proyecto desarrollado, este puede concluir con un entregable sencillo, como la presentación de un reporte final, o incluir más elementos como la implementación y despliegue de una aplicación para ejecutar un modelo y acceder a los resultados en tiempo real. Con base en lo anterior, a continuación, se plantean tres aspectos para tener en cuenta en esta última etapa del proyecto. Po último, se plantea el mantenimiento de productos y resultados entregados como un aspecto opcional posterior a la entrega del proyecto.

2.2.7.1. Despliegue

Al considerar los requerimientos y el alcance del proyecto, el despliegue o la puesta en marcha de la solución desarrollada puede ser llevarse a cabo de manera local, en el equipo del usuario, en servidores locales de la entidad o en la nube. Cabe aclarar que no todos los proyectos requieren este paso; sin embargo, se recomienda trazar un plan de despliegue al inicio, de manera que se puedan estimar los recursos necesarios en tiempo, capacidad de cómputo, personal y demás recursos necesarios para poner marcha de la solución analítica. En caso de no contar con los recursos requeridos, se podrá hacer una planeación más acertada o replantear el proyecto dada su falta.

Por lo general, el ambiente de desarrollo de la solución difiere del ambiente de su producción, debido a lo cual se aconseja el uso de ambientes virtuales o contenedores. Una ventaja para destacar de tales tecnologías es que permiten aislar el ambiente en el que funciona la aplicación, satisfacer las dependencias

requeridas y, a su vez, prevenir daños o conflictos con el sistema operativo que aloja la aplicación.

No solo se debe garantizar el correcto despliegue de la aplicación, también igual manera, es la capacidad de reproducir los modelos implementados y los resultados que estos arrojan. Por lo tanto, se han de aplicar buenas prácticas como utilizar semillas (*seeds*) al llevar a cabo procesos que tengan un componente aleatorio, tanto en el momento preparar los datos y entrenar los modelos, como en el momento de utilizar un modelo entrenado para hacer inferencia.

2.2.7.2. Presentación de resultados

Al presentar los resultados del proyecto se debe tener en cuenta ante quién se da a conocer los resultados y el contexto, ya que esto determina el tipo de enfoque por usar, ya sea técnico, de negocio, financiero u otro. Se aconseja incluir el problema — la necesidad— que origina el proyecto de ciencia de datos, la solución implementada, los *insights* o resultados encontrados, el impacto o beneficio que se genera a partir del proyecto y las limitaciones de la solución.

Al igual que en la etapa de despliegue, la presentación de resultados dependerá del alcance del proyecto y la solución implementada. Los resultados podrán estar contenidos en un informe, una presentación, una capacitación, una reunión de entrega o una combinación de varias de las opciones nombradas.

Además, se recomienda elaborar un reporte del desarrollo del proyecto, en el cual se registren lecciones aprendidas, los enfoques de análisis o las metodologías utilizadas en la solución del problema —tanto las exitosas como las fallidas— y las recomendaciones para situaciones futuras. También, en la medida de lo posible, se recomienda tener una retroalimentación de los actores involucrados en la ejecución del proyecto; así, se podrán identificar oportunidades de mejora para el desarrollo de próximos proyectos.

2.2.7.3. Aprovechamiento del desarrollo realizado

Todo el trabajo llevado a cabo en las etapas y pasos anteriores sería en vano si no se aprovechara de manera correcta. El último paso del proceso consiste precisamente en asegurarse de que los productos, los modelos y las herramientas desarrollados tengan un uso adecuado. Para esto, es preciso que la entidad tenga claro quiénes son los usuarios finales del producto o modelo desarrollado. Esos usuarios finales serán quienes utilicen los resultados del proyecto para apoyar sus labores misionales, entre ellas, el seguimiento a procesos o recursos, la formulación de políticas públicas y la toma de decisiones objetivas basadas en datos.

Es importante que el equipo desarrollador del proyecto se asegure de que estos usuarios finales conozcan los productos entregados, en particular, de que tengan claro los siguientes aspectos:

- Qué se entrega.
- Qué funcionalidades o características ofrece el producto entregado.
- Para qué se puede utilizar el producto entregado; qué se puede lograr con su uso.
- Cómo utilizar el producto entregado.

Si los usuarios finales tienen claros estos aspectos, el producto entregado podrá ser utilizado y aprovechado con éxito aplicando alguna de las varias alternativas que hay, las cuales son complementarias entre sí. En primer lugar, la presentación de los resultados es un buen escenario para socializar los aspectos señalados y asegurarse de que todo quede claro. En segundo lugar, de acuerdo con la complejidad del proyecto y el producto entregado puede ser necesario adelantar sesiones de capacitación por aparte, durante las cuales se explique en detalle todo lo relacionado a los productos entregados. Para terminar, es de suma importancia la documentación del proyecto y de los entregables, por lo cual pueden realizarse varios documentos distintos sobre el proyecto, pero es necesario que se entreguen por lo menos dos:

1

Documento de descripción metodológica.

Informe donde se expliquen aspectos como las fuentes de información utilizadas, el procedimiento llevado a cabo, las técnicas utilizadas y los supuestos que se hicieron durante el desarrollo del proyecto.

2

Documentación sobre las herramientas entregadas.

Por lo general se elabora un manual de usuario que explica cómo instalar y usar la herramienta entregada, y/o cómo interpretar sus resultados.

Finalmente, durante el proceso es inapreciable abrir espacios de retroalimentación para que los usuarios finales expresen lo que piensan sobre los productos entregados. Ese efecto retroactivo de consideraciones posibilita corregir errores, tanto metodológicos como técnicos, que mejore la experiencia de usuario o que se descubran nuevos casos de uso potenciales para las herramientas. Los espacios disponibles pueden ser *presenciales* —en las mismas reuniones de presentación o capacitación—, o *asíncronos*, abriendo canales como encuestas y correos electrónicos para establecer contacto.

Detección y rastreo de inversión pública para el cumplimiento de los ODS

ENTREGA Y APROVECHAMIENTO DEL PRODUCTO FINAL: SEGUIMIENTO A INVERSIÓN PÚBLICA PARA EL CUMPLIMIENTO DE LOS ODS

Para este proyecto se presentaron dos entregables: la base con las 16 medidas de similitud calculadas para los 112.000 proyectos y el tablero de visualización. Estos productos se entregaron a la Comisión ODS del DNP con un manual de usuario, un informe técnico y una presentación para socializar la metodología y los resultados. Es de resaltar que al socializar y divulgar tanto los resultados como los entregables de este proyecto se identificó que sus resultados respondían a una necesidad muy similar de la Dirección de Inversión y Finanzas Públicas (DIFP) del DNP.

2.2.7.4. Mantenimiento de los productos entregados (opcional)

Con las etapas de despliegue, presentación y aprovechamiento de los resultados finaliza la entrega del proyecto de analítica o explotación de datos; sin embargo, según el proyecto es posible que se necesiten gestiones posteriores a su entrega. Una de ellas el mantenimiento de los productos o herramientas entregados, de manera que su desempeño siga siendo el deseado.

Este aspecto se marca como opcional porque depende de la naturaleza del proyecto y de los resultados producidos. Algunos proyectos resultan en un producto puntual, como un informe o un análisis, o una herramienta de propósito único que no cambiará con el tiempo —por ejemplo, un programa que recibe unos datos en un formato estándar y genera unas visualizaciones predefinidas—. En estos casos, probablemente no se requerirá un mantenimiento continuo o la revisión de los productos entregados. En otros casos, sin embargo, la naturaleza de los productos desarrollados podrá demandar una gestión más activa para mantener los resultados de la analítica de datos en correcto funcionamiento.

Las características de productos que pueden requerir mantenimiento posterior a la entrega se describen a continuación.

Actualización de información

Si un producto necesita actualizaciones periódicas de los insumos de información que utiliza, se debe garantizar que siempre cuente con los datos de entrada adecuados. Para esto, deben definirse lineamientos a fin de que la actualización de la información sea correcta; este proceso puede ser manual, en cuyo caso ha de establecerse quién será el encargado de hacerlo, o si es un proceso automático. En el caso de la recolección y actualización automática de datos —por ejemplo, por medio de técnicas de *web scraping*—, puede haber cambios en las tecnologías o configuraciones de los sitios desde donde se trae la información, por lo que será necesario dar mantenimiento al producto para asegurar que los datos sigan estando disponibles. También puede suceder que por razones técnicas, legales o éticas ciertos insumos de información dejen de estar disponibles; en ese caso habrá que efectuar cambios al producto para que pueda funcionar sin esos datos.

Desajuste de modelos

Algunos modelos predictivos pueden desajustarse o descalibrarse con el tiempo, y así empeorar su desempeño como resultado. Por ejemplo, un modelo entrenado con datos actuales para predecir el clima de la próxima semana puede descalibrarse al cabo de un año, cuando las condiciones ambientales sean distintas como consecuencia de cambios climáticos. Si existe el riesgo de desajuste de un modelo, deben definirse procedimientos periódicos de prueba de desempeño de los modelos, para asegurarse de que no desmejore su desempeño con el tiempo. Cuando se identifica que los modelos o productos han sufrido desajuste, se también hay que establecer prácticas para reentrenarlos o reajustarlos, de manera que su desempeño vuelva a ser el deseado.

Productos integrados a sistemas más grandes

Si un producto de analítica está integrado a una plataforma o sistema más grande —por ejemplo, un modelo que analiza el sentimiento de las PQRSD que llegan a una entidad—, debe asegurarse que se mantenga funcional la integración, incluso si se hacen cambios al sistema principal. Para asegurar ello, se requerirá aplicar pruebas funcionales, al igual que eventuales modificaciones al producto de analítica cada vez que se produzcan cambios al sistema principal. También puede haber nuevos requerimientos en protocolos de seguridad de la información y desempeño en tiempos de respuesta en el sistema principal que pueden motivar gestiones de mantenimiento en los productos de analítica.

Los casos mencionados previamente son algunos de los probables escenarios en los cuales puede requerirse el mantenimiento posterior a la entrega de proyectos de analítica y explotación de datos.

Pueden existir otras condiciones posibles, por lo que es fundamental identificar las necesidades de mantenimiento que un proyecto y sus resultados puedan tener.

Lo ideal es hacer esa identificación antes de entregar el proyecto, para que puedan definirse pasos por seguir, lineamientos y responsables de las labores de mantenimiento y seguimiento, y puedan comunicarse en las etapas de presentación de resultados y aprovechamiento de los desarrollos efectuados.

CAPÍTULO 2.3

ECOSISTEMA PARA LA EXPLOTACIÓN Y ANALÍTICA DE DATOS

CUALQUIER ESTRATEGIA O INICIATIVA DE EXPLOTACIÓN DE DATOS QUE DESARROLLEN E IMPLEMENTEN LAS ENTIDADES PÚBLICAS ESTÁ ENMARCADA EN UN ECOSISTEMA PARA LA TRANSFORMACIÓN DIGITAL Y LA ADOPCIÓN DE TECNOLOGÍAS DE LA CUARTA REVOLUCIÓN INDUSTRIAL.

El Gobierno nacional ha implementado, dentro de la política de Gobierno Digital, lineamientos, instrumentos y herramientas que apoyan la adopción del *big data* por parte de las entidades públicas. Parte de estos instrumentos incluyen el Marco de Arquitectura TI, la política de datos abiertos y la integración de los servicios ciudadanos digitales en el Portal Único del Estado colombiano. Este capítulo II.3 describe los actores que forman parte del ecosistema de explotación de datos y propone mecanismos de interacción de las entidades públicas con cada uno de ellos. También se relacionan los recursos —guías metodológicas, modelos, herramientas— con los que cuentan las entidades para mejorar sus capacidades en el aprovechamiento de datos.

2.3.1. MAPEO DE ACTORES DEL ECOSISTEMA

La política de explotación de datos en el país ha visibilizado los esfuerzos del sector privado, la academia, la sociedad civil y la ciudadanía para aprovechar los datos como un activo para la generación de conocimiento e innovación. En este sentido, el desarrollo de la analítica de datos en las entidades públicas demanda de relaciones dinámicas entre distintos actores con el fin de aprovechar al máximo los recursos disponibles, los conocimientos adquiridos y las experiencias que han adelantado para avanzar en la explotación de datos.

Ahora bien, el desarrollo de capacidades en las entidades públicas no es un proceso aislado; por lo contrario, está soportado en una red de actores que integran del ecosistema de explotación de datos en Colombia. El propósito de esta sección es sintetizar las interacciones o alianzas que las entidades públicas pueden establecer con actores del mismo sector público, del sector privado, de la academia y de la sociedad civil para fortalecer las capacidades de explotación de datos y elaborar proyectos de analítica. En la tabla II.3-1 se describe la manera como las entidades del sector público pueden interactuar con los actores del ecosistema de explotación de datos para fortalecer sus capacidades y explorar iniciativas.

Tabla II.3-1. Recomendaciones de articulación de las entidades públicas con actores del ecosistema de explotación de datos

ACTOR	RECOMENDACIONES DE ARTICULACIÓN PARA FORTALECER LA EXPLOTACIÓN DE DATOS EN LAS ENTIDADES PÚBLICAS
Ciudadanos y sociedad civil	• Participar en espacios de *hackatones* organizados por la sociedad civil. Estas actividades ayudan a visibilizar la importancia de los conjuntos de datos publicados por las entidades públicas para solucionar problemáticas de datos. • Establecer canales de diálogo con organizaciones de la sociedad civil para incorporar el enfoque ético y de responsabilidad de tratamiento de datos en el desarrollo de proyectos de analítica de datos.
Gobierno nacional	• Construir una estrecha colaboración con otras entidades públicas para conocer las lecciones aprendidas en el desarrollo de estrategias de *big data* y crear una red de colaboración de intercambio de conocimientos y de *know-how* entre equipos de analítica de datos. Esto permite crear canales más rápidos de aprendizaje para la puesta en marcha de proyectos de analítica y fortalecer las competencias y habilidades del capital humano de las entidades. • Participar en espacios de experimentación y Data-Sandbox dispuestos por el Gobierno para usar las herramientas, técnicas y procesos de analítica de datos de manera exploratoria. • Usar las herramientas y espacios disponibles por el Gobierno nacional en el marco de la política de Gobierno Digital y la Política Nacional de Explotación de Datos.
Sector privado	• Establecer mecanismos de intercambio colaborativo de datos con el sector privado para acceder a datos relevantes que fortalezcan los análisis de política pública; por ejemplo, los datos de telefonía móvil y del sector financiero. Es indispensable anotar que el tratamiento de tales datos debe hacerse dentro del marco normativo de la protección de datos personales y es especialmente esencial la consideración del marco ético para su tratamiento por su carácter sensible. • Es importante definir el propósito claro del intercambio de datos —acuerdos gana-gana— y el objetivo social que se quiere alcanzar con el uso de los datos. Por asuntos de privacidad y seguridad de la información, es recomendable la solicitud de datos agregados o de un intermediario-Data Bróker. Son mecanismos de intercambio de datos con el sector privado las interfaces de programación de aplicaciones API, los acuerdos de colaboración, donaciones de datos, espacios de experimentación, premios, desafíos y otros. • Establecer alianzas con compañías multinacionales que apoyan mediante recursos tecnológicos y financieros el despliegue y la puesta en marcha de proyectos de analítica de datos.

ACTOR	RECOMENDACIONES DE ARTICULACIÓN PARA FORTALECER LA EXPLOTACIÓN DE DATOS EN LAS ENTIDADES PÚBLICAS
Academia	• Participar en espacios académicos, como los foros o los *workshops* para compartir, experiencias, técnicas, conocimiento en analítica de datos. • Explorar alianzas con universidades y centros de excelencia en *big data* con el propósito de apoyar el uso de las tecnologías de análisis de grandes conjuntos de datos y Data Analitycs, para la formación de capital humano y para el desarrollo de la investigación en la entidad.
Centros de investigación e innovación	• Gestionar alianzas con centros de investigación e innovación que comúnmente sirven como articuladores entre sector privado, academia, Gobierno y startups para impulsar retos y desafíos que puedan ser solucionados por los actores especializados del ecosistema de datos.
Agencias de cooperación internacional	• Gestionar desde la alta dirección de las entidades, alianzas con las agencias de cooperación internacional para impulsar el fortalecimiento de capacidades en las entidades. Lo anterior dado que el Banco Interamericano de Desarrollo, CAF, el Centro Económico para el Desarrollo de América Latina y las Naciones Unidas y el Banco Mundial apoyan desde diferentes iniciativas el reconocimiento y acompañamiento de proyectos que mejoren el ciclo de las políticas públicas y generen valor público a partir del uso de los datos.

Fuente: elaboración propia.

2.3.2. RECURSOS DISPONIBLES PARA EL FORTALECIMIENTO DE CAPACIDADES EN LAS ENTIDADES

Para fortalecer las capacidades habilitantes para la explotación y analítica de datos, las entidades públicas no requieren iniciar el camino desde cero. Para cada una de las capacidades nacionales se dispone de recursos o herramientas destinadas a apoyar la estrategia de explotación de datos. En las tablas siguientes se incluye el mapeo de herramientas disponibles para que las entidades públicas las incorporen en su estrategia de explotación de datos.[9]

9. Los enlaces que se incluyen en las tablas pueden ser dinámicos. Por lo que se recomienda al momento de su consulta en línea, confirmar los títulos de los documentos.

Tabla II.3-2. Recursos para el fortalecimiento del recurso humano para la explotación de datos

REQUERIMIENTO	RECURSO
Capacitación y formación de talento humano	• Plan Nacional de Formación y Capacitación 2020-2030 • Eje de Transformación Digital: temáticas definidas para mejorar las competencias de los funcionarios y servidores públicos en analítica de datos, ética en el uso de los datos, gestión de datos, entre otros. https://www.funcionpublica.gov.co/eva/admon/files/empresas/ZW1wcmVzYV83Ng==/imgproductos/1450185065_2ef719ee0eb3b2141b1a7e53bb98b887.pdf • Plan de talento digital de MinTIC: Fomentar las capacidades y habilidades digitales de los ciudadanos- Consulta de convocatorias: Convocatorias (mintic.gov.co) • Estrategia de uso y apropiación de Arquitectura TI del MinTIC. Incluye jornadas de capacitación e interacción con expertos. • Jornadas de capacitación para el uso, reutilización y publicación de datos abiertos del MinTIC.

Fuente: elaboración propia.

REQUERIMIENTO	RECURSO
Adquisición de servicios profesionales	Acuerdo Marco de Nube Pública que permite adquirir la prestación de servicios de diferentes perfiles junio y senior entre los que se encuentran: Validación de datos, muestras de datos, integración de bases. • Arquitecto de nube pública • Experto en almacén de datos • Experto en bases de datos • Experto en *big data* • Experto en inteligencia de negocios • Experto en servidores y aplicación web

Fuente: elaboración propia.

Tabla II.3-3. Recursos para el fortalecimiento de las capacidades tecnológicas

REQUERIMIENTO	RECURSO
Infraestructura para el almacenamiento y procesamiento de datos	Acuerdo Marco de Nube Pública y Nube Privada para adquirir servicios de computación en la nube ofrecidos por diferentes proveedores. A través de esto es posible adquirir infraestructura de servidores, procesamiento y almacenamiento, bases de datos, contenedores, servicios de Inteligencia Artificial y de big data, Internet de las cosas y Data SandBox https://www.colombiacompra.gov.co/tienda-virtual-del-estado-colombiano/tecnologia/nube-publica-iii
	Herramientas de código abierto para el desarrollo de iniciativas de analítica de datos https://www.softwarepublicocolombia.gov.co/es/public-software

Fuente: elaboración propia.

REQUERIMIENTO	RECURSO
Arquitectura TI para la interoperabilidad	• Marco de interoperabilidad de la Política de Gobierno Digital http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_gobierno_digital.pdf • Modelo de madurez del marco de interoperabilidad http://lenguaje.mintic.gov.co/sites/default/files/archivos/marco_de_interoperabilidad_para_gobierno_digital.pdf

Fuente: elaboración propia.

REQUERIMIENTO	RECURSO
Elaboración del Plan Estratégico de Tecnologías de la Información	• Herramienta para la construcción del PETI https://www.mintic.gov.co/arquitecturati/630/w3-article-15031.html • Guía para la construcción del PETI - Gobierno Digital https://www.mintic.gov.co/arquitecturati/630/w3-article-15031.html
Construcción de arquitectura empresarial	• Guía general para un proceso de Arquitectura empresarial https://www.mintic.gov.co/arquitecturati/630/articles-9435_Guia_Proceso.pdf

Fuente: elaboración propia.

Tabla II.3-4. Recursos para el fortalecimiento de las capacidades organizacionales

REQUERIMIENTO	RECURSO
Marco de política para la explotación de datos y marco jurídico aplicable a la explotación de datos	1. Política Nacional de Explotación de Datos - Documento CONPES 3920-2018 https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3920.pdf 2. Política Nacional para la Transformación Digital y la Inteligencia Artificial - Documento CONPES 3975 de 2019 https://www.mintic.gov.co/portal/604/articles-107147_recurso_1.pdf 3. Marco regulatorio para protección de datos personales e intercambio de datos entre entidades públicas. 4. Lineamientos de la Superintendencia de Industria y Comercio para la aplicación del principio de responsabilidad demostrada, la implementación de la privacidad y la ética desde el diseño y el defecto https://www.sic.gov.co/noticias/guia-para-la-implementacion-del-principio-de-responsabilidad-demostrada 5. Guía de la Gestión ética de los datos-Banco Interamericano de Desarrollo. 2019 https://publications.iadb.org/publications/spanish/document/La_Gesti%C3%B3n_%C3%89tica_de_los_Datos.pdf 6. Marco ético de inteligencia artificial - DAPRE https://dapre.presidencia.gov.co/AtencionCiudadana/convocatorias-consultas/consulta-200813-marco-ia-colombia

REQUERIMIENTO	RECURSO
Gestión y operativización del ciclo de vida de los datos	1. Guías de lineamientos de Arquitectura TI dispuestas en la Política de Gobierno Digital de MinTIC • Guía técnica básica de información • Guía de administración de dato maestro • Guía del ciclo de vida del dato • Guía de migración del dato • Guía para construir el catálogo de componentes de información 2. Guía de uso y aprovechamiento de datos abiertos en Colombia https://herramientas.datos.gov.co/sites/default/files/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf 3. Guía de estándares de calidad e interoperabilidad de los datos abiertos del gobierno de Colombia https://herramientas.datos.gov.co/sites/default/files/A_guia_de_estandares_final_0.pdf 4. Guía de transformación de datos abiertos a datos enlazados https://herramientas.datos.gov.co/sites/default/files/Guia%20Linked%20Open%20Data.pdf 5. Guía de anonimización de datos estructurados (Archivo General de la Nación) https://www.archivogeneral.gov.co/consulte/recursos/publicaciones 6. Guía de herramientas de analítica para la explotación de datos https://herramientas.datos.gov.co/sites/default/files/Inventario%20herramientas%20anal%C3%ADtica_0.pdf

Fuente: elaboración propia.

2.3.3. DIAGNÓSTICO DEL NIVEL DE CAPACIDADES QUE TIENE LA ENTIDAD PARA AVANZAR EN LA EXPLOTACIÓN DE DATOS

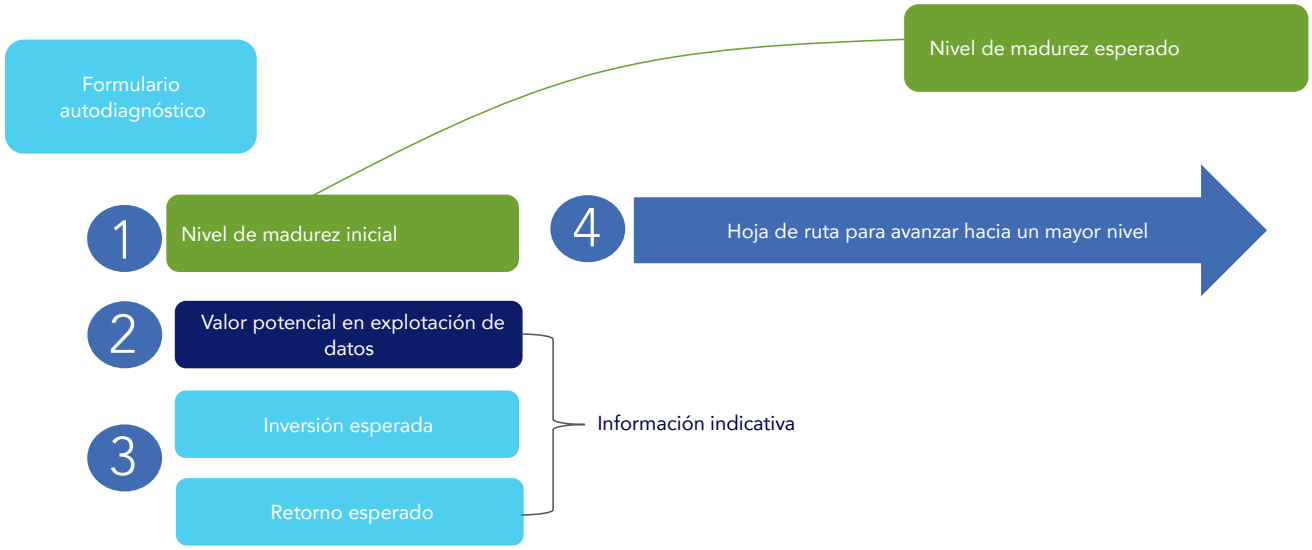
El Modelo de explotación de datos para las entidades públicas de Colombia es una herramienta diseñada por la Dirección de Desarrollo Digital del DNP en conjunto con Infométrika[10], que permite a las entidades públicas nacionales y territoriales hacer un diagnóstico del nivel de madurez en explotación de datos de acuerdo con seis dimensiones:

El modelo permite, además, que las entidades identifiquen el nivel de madurez en explotación de datos que desean alcanzar con base en su misionalidad, y disponer de puntos de referencia para construir una hoja de ruta que les permita avanzar en el fortalecimiento de sus capacidades para la explotación de datos (figura II.3-1).

Este modelo permite que las entidades tengan recomendaciones para estimar las necesidades de inversión en recurso humano y tecnológico que deben prever para alcanzar el nivel de madurez en explotación de datos que desean alcanzar. También permite estimar cuál sería el valor potencial que podrían alcanzar las entidades a partir de la inversión en explotación de datos. Lo anterior, teniendo como referencia el valor público que genera la entidad a partir de la explotación de datos y *big data*.

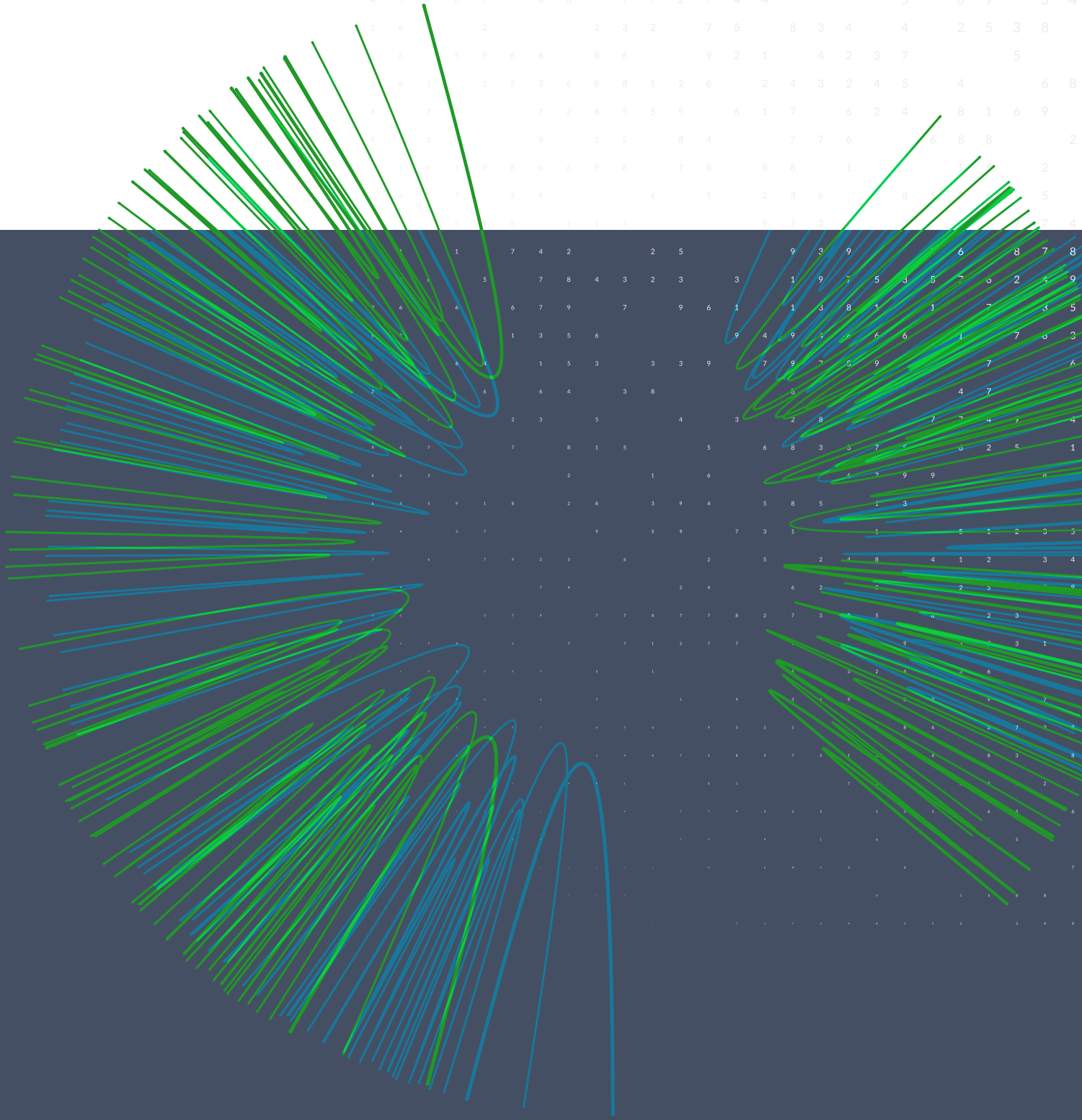
10. Este modelo se en el marco del contrato de consultoría 658 de 2020 celebrado entre el Departamento Nacional de Planeación y Consultores en Información Infométrika SAS. Las dimensiones del modelo de madurez de *big data* se tuvieron en cuenta para el planteamiento de las capacidades de las entidades, mencionadas en el capítulo I de la parte II de la presente publicación. Por lo tanto, algunas de las preguntas orientadoras mencionadas en ese capítulo, corresponden a preguntas del formulario diagnóstico que instrumentaliza el modelo de explotación de datos elaborado por el DNP e Infométrika.

Figura II.3-1. Modelo de explotación de datos para las entidades públicas en Colombia



Fuente: elaboración propia.

Para la instrumentalización del modelo se cuenta con un formulario de diagnóstico para que las entidades públicas evalúen su nivel de madurez en *big data*. El formulario también posibilita identificar el efecto en la generación de valor público a partir de la explotación de datos. A la vez, el modelo incorpora una hoja de recomendaciones estándar para que las entidades mejoren sus capacidades para la explotación de datos de acuerdo con el nivel que desean alcanzar. El modelo de implementación de explotación de datos está integrado en el *Manual de la Política de Gobierno Digital*, una herramienta que les faculta a las entidades públicas alcanzar uno de los propósitos de la Política de Gobierno Digital: la toma de decisiones basadas en datos.



PARTE 03

Casos de aprovechamiento de datos del sector público en el contexto de la COVID-19.

Experiencia manos en la data

INTRODUCCIÓN

LA CRISIS OCASIONADA POR LA COVID-19 HA VISIBILIZADO LA IMPORTANCIA DE LOS DATOS PARA RESOLVER PROBLEMÁTICAS DE POLÍTICA PÚBLICA EN MATERIA DE SALUD PÚBLICA, MOVILIDAD, EDUCACIÓN, Y REACTIVACIÓN ECONÓMICA. ASIMISMO, LA ACTUAL CRISIS PUSO DE MANIFIESTO LA NECESIDAD DE ARTICULAR ESFUERZOS ENTRE LOS SECTORES PÚBLICO, PRIVADO Y LA ACADEMIA PARA AUMENTAR LA DISPONIBILIDAD Y EL APROVECHAMIENTO DE DATOS, EN EL MARCO DE LA NORMATIVA APLICABLE A LA PROTECCIÓN DE DATOS PERSONALES, LA PRIVACIDAD DE LAS PERSONAS Y LA PROPIEDAD INTELECTUAL.

El Departamento Nacional de Planeación como entidad líder en la coordinación, articulación y apoyo de la planificación y orientación del ciclo de políticas públicas del Gobierno nacional, ha promovido el diseño de un marco de política que aumente el aprovechamiento de los datos para tomar decisiones y para generar valor social y económico. El CONPES 3920: *Política Nacional de Explotación de Datos (big data)*, el Documento CONPES 3975: *Política Nacional para la Transformación Digital e Inteligencia Artificial* y el CONPES 4023: *Política para la reactivación, la repotenciación y el crecimiento sostenible e incluyente: nuevo compromiso por el futuro de Colombia*, han constituido las bases para continuar consolidando la infraestructura de datos del país, y las condiciones habilitantes para aumentar su aprovechamiento.

Uno de los ejes de acción del marco de política para explotar los datos en el sector público y generar un efecto positivo en la generación de valor público es aumentar la cultura de datos en las entidades públicas, así como las capacidades del capital humano para aprovechar los datos y resolver asuntos de interés público. Por lo anterior, el Departamento Nacional de Planeación se sumó a la iniciativa de Manos en la Data (MeD), promovida por CAF, cuyo objetivo consiste en propiciar el uso de datos intensivo, eficiente y seguro dentro del sector público en países de América Latina.

Manos en la Data (MeD) incluye una metodología de trabajo para la producción de prototipos de ciencia de datos que atiendan una problemática o pregunta de política pública muy concreta, y lo hagan de manera rápida, colaborativa y costo-efectiva. Esa metodología se aplica en cinco etapas, son ellas: la selección de preguntas por resolver, la conformación de equipos de trabajo para desarrollar los prototipos que respondan a las preguntas planteadas, un *workshop* que hace las veces de *kick-off* para cada proyecto, la etapa de desarrollo —con un mínimo de ocho semanas—, y la entrega final de los prototipos que incluye capacitaciones *hands-on* para los funcionarios participantes.

Cada edición de MeD se organiza como una alianza de tres partes claves: una agencia pública con rol preponderante en el uso de información para el diseño e implementación de políticas públicas, una institución de vinculación científico-tecnológica y la Vicepresidencia de Conocimiento de CAF, que a través de un grupo de investigadores coordina y monitorea el trabajo de los equipos conformados por funcionarios y científicos de datos. El DNP integró la alianza junto con CAF y Alianza CAOBA para la realización de Manos en la Data-Colombia, que fue la tercera réplica de esta iniciativa, antes desarrollada en Argentina y Uruguay.

Además de su función de coordinación en conjunto con los investigadores de CAF, el DNP aportó en la identificación de problemáticas de interés que pudieran ser resueltas a partir de proyectos de analítica de datos, en el marco de la COVID-19. Y a la vez, desde el DNP se contó con la participación permanente de los asesores de las direcciones técnicas expertas en cada una de las temáticas de política pública abordadas. Por otro lado, la participación de Alianza CAOBA, como centro de excelencia para la puesta en marcha de los proyectos de analítica de datos, aportó los equipos de científicos de datos que lideraron el desarrollo de los prototipos para atender a las preguntas de política pública priorizadas para Manos en la Data.

CAF y el DNP se complacen en presentar a través de esta publicación los resultados de los proyectos patrocinados en el marco de la iniciativa Manos en la Data - Colombia en la que se desarrollaron seis proyectos de analítica de datos. Cabe destacar que todos los proyectos se enmarcaron en la coyuntura de la COVID-19 y, por lo tanto, constituyen un claro ejemplo del modo como el aprovechamiento de datos aporta a la comprensión y atención de problemáticas públicas, aun en casos de emergencia, tal como la que planteó la pandemia en el año 2020. Los proyectos tratan temas muy variados, que van desde logística y transporte de carga carretero, provisión de servicios de cuidados en la órbita de la salud, seguridad ciudadana, nutrición y alimentación de niñas y niños en riesgo de desnutrición, hasta estrategias para la recuperación económica y el monitoreo del empleo formal en el país.

RESULTADOS DE LOS PROYECTOS DE MANOS EN LA DATA - COLOMBIA

En esta sección se presenta la descripción y resultados de los seis proyectos de analítica de datos. La elaboración de cada proyecto contó con la participación de los asesores de las direcciones técnicas del DNP y del equipo de científicos de datos de Alianza CAOBA. Cada equipo contó con aproximadamente dos meses para la elaboración del proyecto. El resumen de cada proyecto presentado en los apartados siguientes fue tomado de los reportes finales elaborados por los integrantes de los proyectos de Manos en la Data - Colombia.

CAPÍTULO 3.1

CARACTERIZACIÓN DE ZONAS DE CONCENTRACIÓN DE LA COVID-19 POR MUNICIPIOS EN EL MARCO DE LA REACTIVACIÓN ECONÓMICA EN COLOMBIA[11]

3.1.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

La presencia y evolución de la COVID-19 ha requerido de la gestión púbica de los Gobiernos locales del país en aspectos sociales, económicos y productivos. En Colombia se evidencia actualmente una disparidad de realidades en el territorio nacional relacionadas con el manejo que se le ha dado a la pandemia. Tales diferencias pueden estar dadas por factores propios de las entidades territoriales, así como por los mecanismos de control implementados. El diseño e implementación de estrategias de reactivación en los municipios requiere de conocer principalmente tres dinámicas:

1. La económica
2. La de contagio de la COVID-19
3. La dinámica de movilidad con municipios socios estratégicos.

La comprensión de las características identificadas será fundamental para la construcción de hojas de ruta de atención a la crisis en los territorios.

11. Integrantes del equipo: José Ramón Romero Pineda, Luz Karine Ardila Vargas, David Alejandro Huertas, Catalina Alvarado Rojas, Manuel Alejandro Castañeda García, Wilson Javier Arenas López.

3.1.2. OBJETIVOS DEL PROYECTO

Objetivo general

Caracterizar el estado y el manejo de la pandemia por las entidades territoriales.

Objetivos específicos

Identificar zonas de contagio de la COVID-19 por municipio para todo el territorio nacional.

Clasificar a los municipios según su estado y su manejo de la pandemia.

Desarrollar una herramienta de apoyo en el proceso de reactivación y manejo de la pandemia.

3.1.3. DATOS UTILIZADOS

Las bases de datos utilizadas en este proyecto fueron en su mayoría datos abiertos y datos generados en la Dirección de Descentralización y Desarrollo Regional (DDDR) del DNP. Los datos se clasificaron de la siguiente forma: datos sociodemográficos, datos de movilidad, datos epidemiológicos, datos económicos. El detalle de cada una de las fuentes de datos se muestra en la tabla III.1-1.

Tabla III.1-1. Fuentes de datos utilizadas

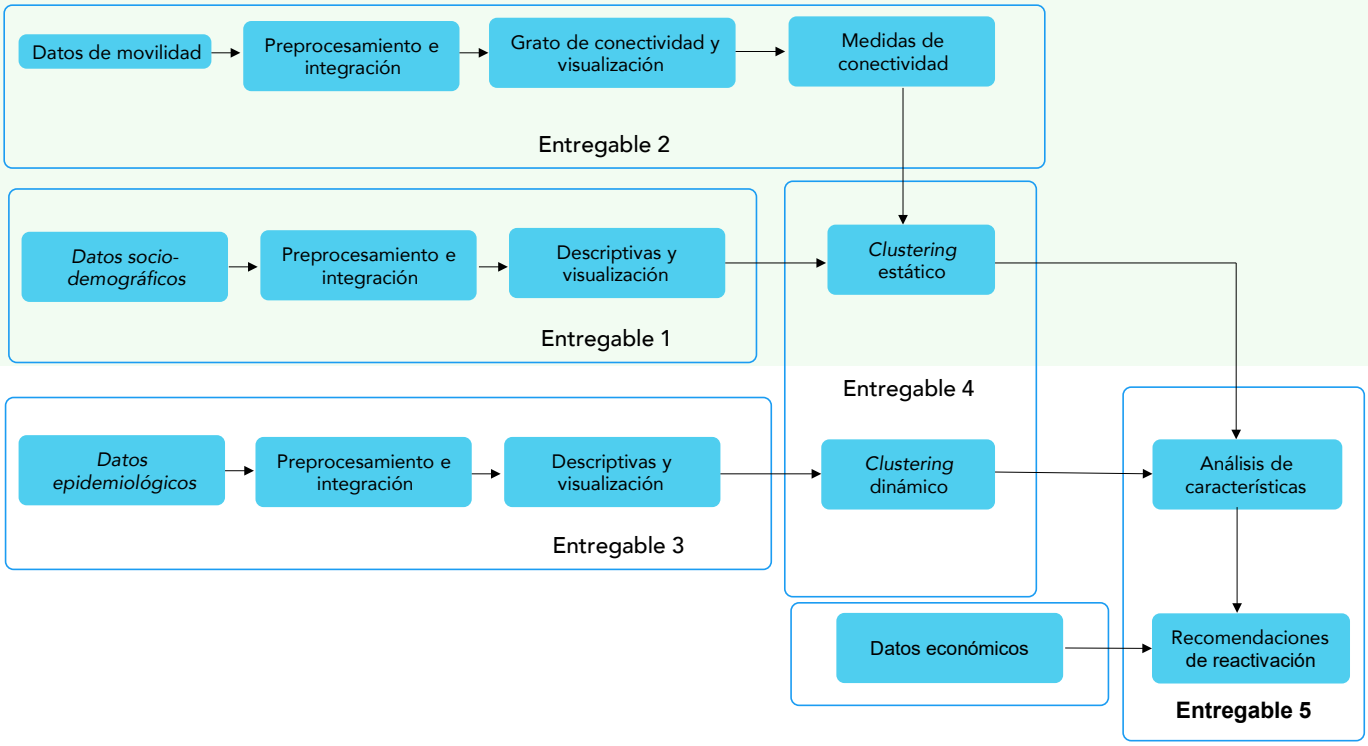
CATEGORÍA	FUENTE	DESCRIPCIÓN
Datos sociodemográficos	TerriData	Densidad poblacional
	Censo	Población - Densidad poblacional - Hacinamiento - IPM - Población por rangos decenales - Problemas de salud - Acceso a agua - Autorreconocimiento étnico
Datos de movilidad	Matriz de subregiones funcionales	Matriz de viajes terrestres origen/ destino por actividad entre subregiones funcionales a nivel nacional
	Vuelos nacionales	Matriz de viajes aéreos origen/ destino a nivel nacional
Datos epidemiológicos	Boletín epidemiológico	Número de casos diarios de la COVID-19: localización, estado, tipo
	Datos abiertos	Número de casos diarios de la COVID-19 por municipio
Datos económicos	FUT	Ingreso tributario y no tributario
	Estadísticas nacionales	VA per cápita, proyecciones de población

Fuente: elaboración propia.

3.1.4. MODELO PROPUESTO

A partir de los datos presentados en la sección III.1.3, se realizaron distintos análisis y su respectiva visualización, con el fin de plantear indicadores para que las entidades territoriales los utilicen para conocer el estado actual de la pandemia —contagios, recuperados, pruebas—, posibles características sociodemográficas y de movilidad asociadas a ese estado, y probables efectos de esas características durante la reactivación económica. El procesamiento de cada conjunto de datos inició con la preparación de los datos, la revisión de las fuentes de información y la selección de variables relevantes para el estudio. Adicionalmente se definió la resolución espacial de cada fuente de información y la temporalidad de los datos frente a si eran estáticos o dinámicos El diagrama de bloques de la solución propuesta se puede observar en la figura III.1-1.

Figura III.1-1. Diagrama de bloques de la solución



Fuente: tomado de informe final (DNP & CAOBA, 2020)

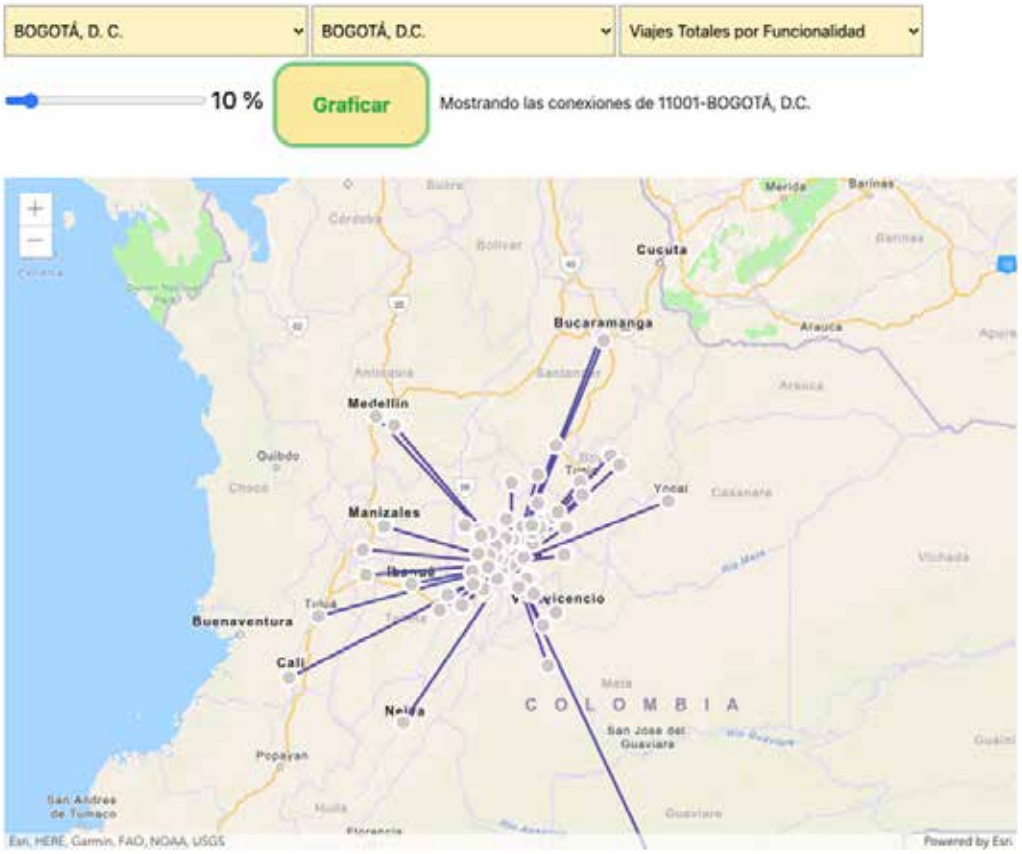
3.1.5. RESULTADOS

De los diferentes análisis de datos implementados, se obtuvieron seis principales resultados, que se integraron en un tablero de control y se publicaron en el portal del DNP. A continuación, se describen los resultados de la visualización.

3.1.5.1. Conexiones por funcionalidades (Grafo de conectividad)

Este tablero permite visualizar la conectividad entre diferentes municipios del país teniendo en consideración la intensidad de los desplazamientos terrestres y aéreos entre municipios desagregado por motivo de desplazamiento (por salud, por trabajo, por educación). El tablero permite seleccionar por municipio cuales son los 10 municipios aliados, con el objetivo de que los tomadores de decisiones puedan coordinar acciones para la mitigación de riesgos.

Figura III.1-2. Tablero de visualización de conexiones por funcionalidades

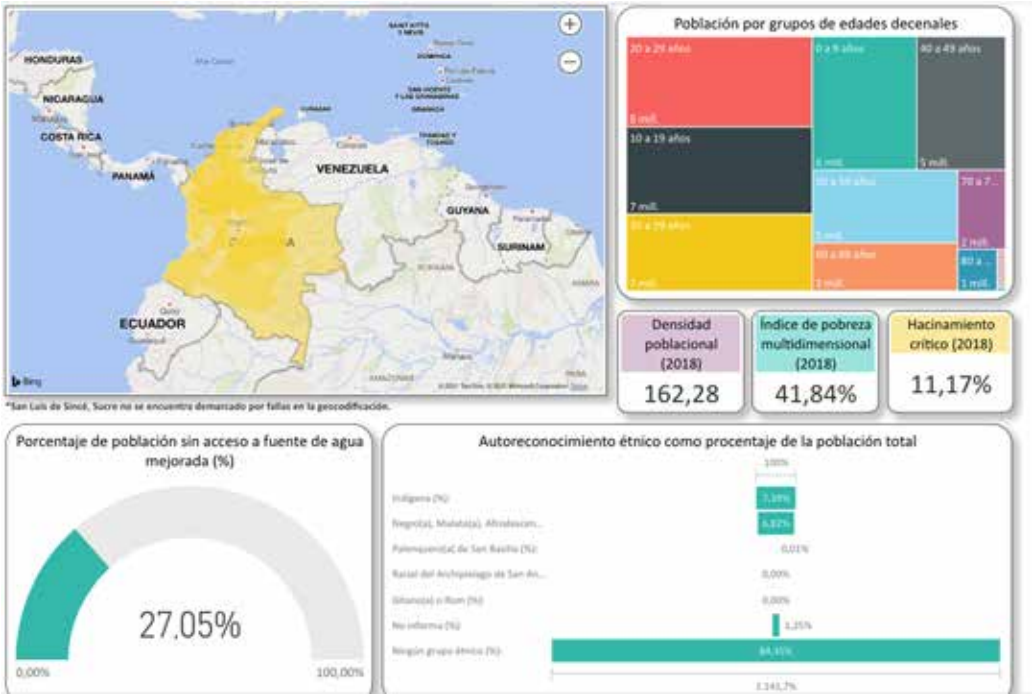


Fuente: tomado de informe final (DNP & CAOBA, 2020)

3.1.5.2. Estadísticas sociodemográficas

El tablero de control igualmente incluye la visualización de características sociodemográficas de los municipios. El usuario tiene la opción de elegir el municipio sobre el cual desea consultar y se muestra la población según edades decenales, el IPM y el índice de hacinamiento crítico. Además, se muestra el porcentaje de población que reportó algún problema de salud sin hospitalización y el autorreconocimiento étnico de la población total (DNP & CAOBA, 2020).

Figura III.1-3. Tablero de visualización de estadísticas sociodemográficas



Fuente: tomado de informe final (DNP & CAOBA, 2020)

3.1.5.3. Agrupamiento sociodemográfico (*Clustering* estático)

Las variables sociodemográficas permitieron el diseño de un algoritmo de *clustering* que permitió identificar cinco grupos de municipios, según su vulnerabilidad sociodemográfica. Las variables de entrada para la visualización del clúster estático son las siguientes: densidad poblacional, proporción de grupos etarios, conformación de los hogares, prevalencia de hipertensión, diabetes, cáncer, hacinamiento, índice de pobreza multidimensional, ingreso tributario y no tributario.

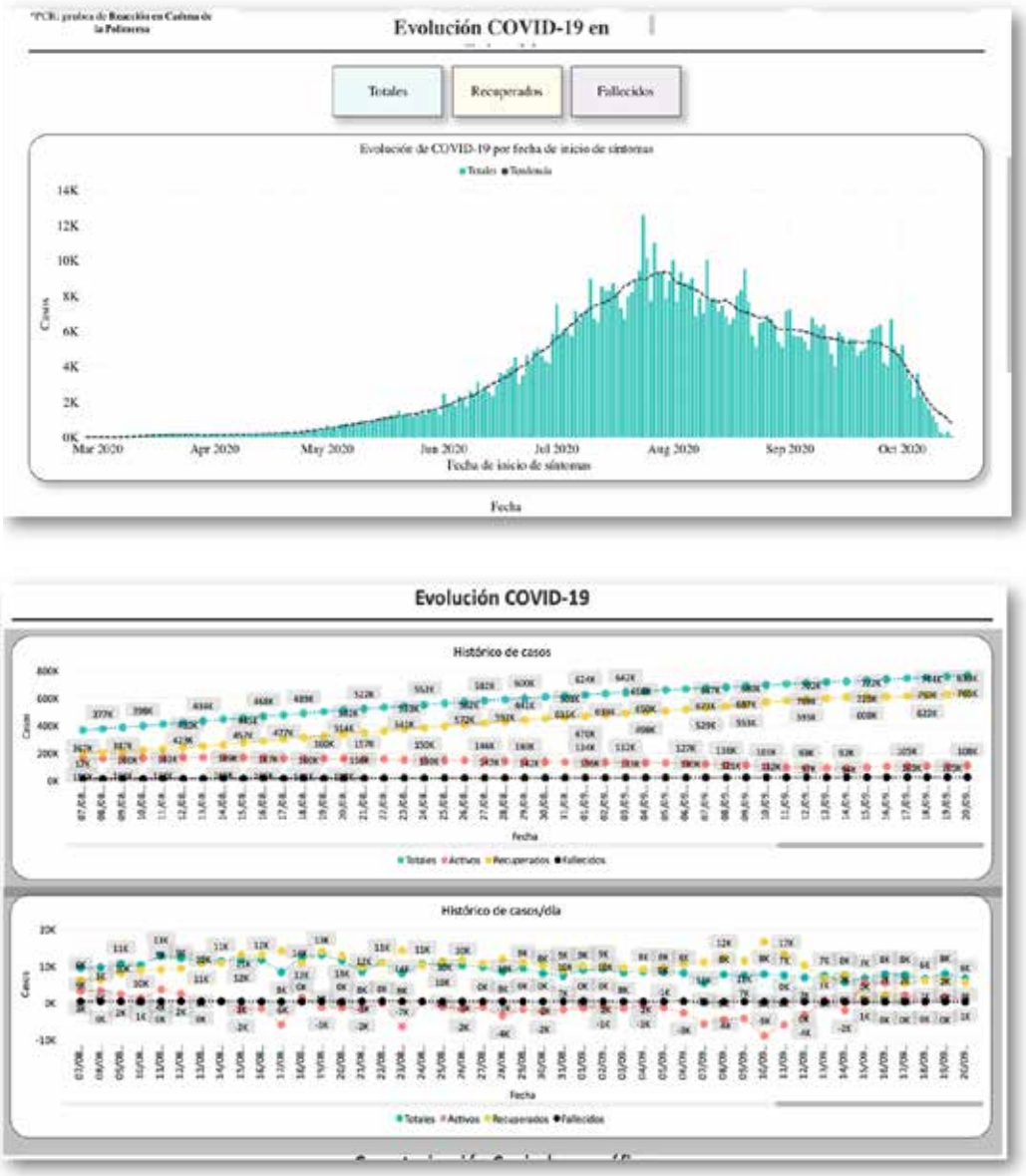
Figura III.1-4. Tablero de visualización de *clúster* estático

Fuente: tomado de informe final (DNP & CAOBA, 2020)

3.1.5.4. Estadísticas de la COVID-19 (Caracterización epidemiológica)

Este tablero de visualización contiene variables epidemiológicas que permiten evidenciar el estado actual de la pandemia en determinado municipio. El tablero habilita la consulta por departamento, municipio, y fecha a partir de la cual se desean ver los datos. La visualización contiene los casos totales, casos activos, casos recuperados y fallecidos. También permite la visualización de estadísticas como número de casos por millón de habitantes, tasas de recuperación y mortalidad; distribución porcentual por departamento y por municipio, por tipo de recuperación, etnia, estado de casos, tipo de casos y género.

Figura III.1-5. Tablero de visualización caracterización epidemiológica

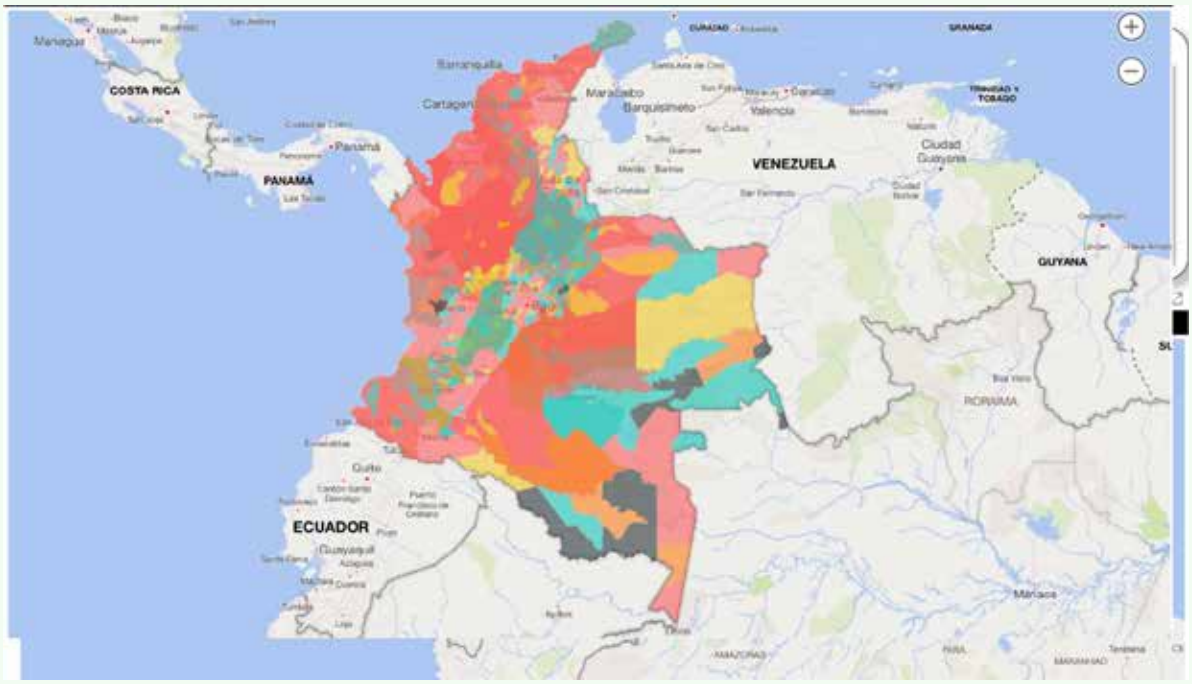


Fuente: tomado de informe final (DNP & CAOBA, 2020)

3.1.5.5. Agrupamiento de la COVID-19 (*Clustering* dinámico)

Se elaboró un algoritmo de *clustering* dinámico, a partir de las variables epidemiológicas. En este análisis se definieron cinco grupos de municipios según su nivel de vulnerabilidad ante la pandemia. Los resultados de ese análisis se pueden visualizar en color en el mapa de Colombia.

Figura III.1-6. Tablero de visualización *clustering* dinámico - Grupos de municipios



Fuente: tomado de informe final (DNP & CAOBA, 2020)

3.1.5.6. Caracterización económica y reactivación

Se diseñó un tablero de control para visualizar variables relevantes en la reactivación económica y el manejo de la pandemia, en cuyos paneles se muestra información de la estructura económica y productiva del municipio o el departamento, la distribución del número de establecimientos formales por actividad económica, la distribución del valor agregado per cápita según sector económico. Este tablero tiene por objetivo apoyar la toma de decisiones en territorio brindando un panorama general de su economía en el marco de la reactivación.

Figura III.1-7. Tablero de visualización clustering dinámico - Panorama general

Fuente: tomado de informe final (DNP & CAOBA, 2020).

3.1.5.7. Conclusiones

Los tableros elaborados en este proyecto aportan un análisis pertinente para los tomadores de decisiones, para contribuir al diseño de las medidas de reactivación territorial y se diferencian de otros tableros previamente desarrollados en el país. El tablero elaborado por DNP y CAOBA en el marco de MeD, aplicó métodos analíticos sobre dichas características, particularmente *clustering* y medidas de redes y grafos (DNP & CAOBA, 2020). Adicionalmente, este proyecto se abordó desde un enfoque subregional en apoyo con los datos de la matriz de subregiones funcionales y la orientación del equipo de políticas públicas del DNP.

CAPÍTULO 3.2

PROYECTO ACTUALIZACIÓN DE LA MATRIZ ORIGEN-DESTINO DE TRANSPORTE DE CARGA EN MODO CARRETERO[12]

3.2.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

La crisis social y económica de la COVID-19 ha afectado de manera negativa la dinámica del transporte intermunicipal e interdepartamental en Colombia. Por lo anterior surge la necesidad de identificar la dinámica actual de la carga carretera en el país provocada por la COVID, analizar el impacto de la COVID-19 en esta actividad económica y contar con insumos para aportar a los tomadores de decisiones relacionados con la materia.

3.2.2. OBJETIVOS DEL PROYECTO

Objetivo general

Actualizar la *matriz origen-destino* que combine todas las posibles duplas origen-destino, con una zonificación municipal del territorio nacional y cuyas dimensiones correspondan a las siguientes variables: volumen transportado en kilogramos, volumen transportado en galones, cantidad de viajes totales, cantidad de viajes con valor a cero (viajes sin carga), cantidad de viajes líquidos, distancia en kilómetros y tiempo de viaje en minutos.

Objetivos específicos

- Conocer cuál ha sido el impacto en el transporte de carga modo carretero de la emergencia sanitaria por la COVID-19, en términos de las variables de volumen en kilogramos, volumen en galones y cantidad de viajes totales realizados mensualmente.
- Analizar el impacto desagregado por pares origen-destino por municipios, enfocado en la afectación por la emergencia declarada como respuesta a pandemia causada por la COVID-19.

12. Integrantes del equipo: Edwin Montoya, Diana Carolina Benjumea Hernández, Sebastián Ospina Porras, Óscar Andrés Patiño Putumayo, Cesar Moreno Boyacá, Daniel Gaitán Forero.

3.2.3. DATOS UTILIZADOS

Para la estructuración de la *matriz origen-destino* se tuvieron en cuenta los conjuntos de datos especificados en la tabla 2-1, clasificados en las siguientes categorías: distancia de viajes, tiempo de viajes, registros de viajes e información geográfica.

Tabla III.2-1. Fuentes de datos utilizadas

CATEGORÍA	FUENTE	DESCRIPCIÓN
Distancia de viajes	Matriz origen destino suministrada por API de ArcGIS	Datos de distancia en kilómetros entre los pares origen-destino zonificados por municipio.
Tiempo de viajes	Matriz de origen-destino suministrada por el Ministerio de Transporte	Matriz origen-destino cuya única dimensión es el tiempo de viaje en minutos en pares origen-destino zonificados por municipio.
Registros de viajes	Registro Nacional de Despachos de Carga	Responde al consolidado de registros de viajes de transporte de carga modo carretero en el periodo enero de 2015 a agosto de 2020.
Información geográfica	Diccionario de datos suministrado por el Ministerio de Transporte	Diccionario de variables que contiene los códigos numéricos, nombres de los lugares, municipios y departamentos de Colombia.

Fuente: elaboración propia.

3.2.4. MODELOS Y PRINCIPALES RESULTADOS

El primer resultado del proyecto es la actualización de la matriz origen destino a partir de la consolidación de los datos históricos de transporte de carga en modo carretero en Colombia. La matriz automatiza la generación de una matriz de pares origen-destino y está constituida por siete variables.

Tabla III.2-2. Variables de la matriz origen destino

VARIABLES
Cantidad total de volumen en kilogramos trasportado entre municipios. Este dato responde a la carga sólida trasportada.
Cantidad total de volumen en galones trasportado entre municipios. Este dato responde a la carga líquida trasportada.
Cantidad total de viajes realizados sin carga (viajes en vacío).
Cantidad total de viajes realizados con carga líquida.
Distancia promedio en kilómetros estimada mediante ArcGIS entre cada par origen-destino.
Tiempo promedio de viaje en minutos entre cada par origen-destino.

Fuente: elaboración propia.

Para la visualización de la *matriz origen-destino* se construyó un tablero de visualización en PowerBI (figura III.2-1), formado por siete hojas que contienen gráficas y tablas que incluyen diferentes dimensiones sobre las cuales es posible analizar la información. El tablero permite generar información histórica, obtener cifras por departamento, por características de la carga, cifras —acumuladas o mensualizadas— y anuales.

Figura III.2-1. Tablero de visualización matriz origen-destino

Departamento Origen	ANTIOQUIA	ARAUCA	ATLÁNTICO	BOGOTÁ. D.C.	BOLÍVAR	BOYACÁ	CALDAS	CAQUETÁ	CASANARE	CAUCA	CESAR	CHOCÓ	CÓRDOBA
ANTIOQUIA	43,72%	0,04%	4,79%	7,89%	6,28%	1,20%	2,23%	0,13%	0,13%	0,68%	0,56%	0,59%	1,97%
ARAUCA	0,10%	52,18%	0,20%	9,62%	0,18%	3,02%	0,01%		24,78%		0,07%		0,00%
ATLÁNTICO	15,28%	0,06%	9,14%	8,26%	12,46%	1,61%	0,67%	0,22%	0,47%	0,44%	5,77%	0,04%	5,45%
BOGOTÁ. D.C.	12,05%	0,46%	6,83%	0,14%	5,72%	7,73%	1,73%	0,55%	1,50%	0,86%	0,74%	0,07%	0,65%
BOLÍVAR	20,66%	0,13%	20,32%	15,08%	4,54%	1,62%	0,73%	0,02%	0,44%	0,76%	3,27%	0,03%	2,60%
BOYACÁ	5,11%	0,72%	7,40%	10,27%	5,59%	17,74%	0,45%	0,12%	2,23%	0,50%	0,28%	0,04%	0,15%
CALDAS	18,86%	0,01%	4,03%	11,96%	7,15%	1,93%	6,38%	0,20%	0,22%	1,36%	0,38%	0,38%	0,91%
CAQUETÁ	2,38%	0,01%	0,08%	17,48%	0,80%	4,29%	0,20%	16,38%	0,42%	1,58%	5,85%		0,00%
CASANARE	1,46%	0,84%	5,77%	11,97%	1,87%	3,39%	0,12%	0,04%	44,76%	0,05%	0,70%		0,62%
CAUCA	8,57%	0,18%	4,10%	9,79%	2,18%	0,55%	1,42%	0,19%	0,25%	16,15%	0,55%	0,05%	0,43%
CESAR	4,33%	0,03%	22,28%	4,30%	6,61%	3,51%	0,56%	0,02%	0,51%	0,03%	9,84%	0,01%	0,69%
CHOCÓ	33,68%		0,78%	5,05%	2,29%	0,03%	7,36%		0,25%	0,39%	0,06%	0,92%	0,70%
CÓRDOBA	16,81%	0,00%	18,59%	3,06%	23,46%	0,87%	0,76%	0,01%	0,13%	3,32%	4,50%	0,02%	8,74%
CUNDINAMARCA	7,42%	0,14%	8,64%	18,64%	3,48%	6,13%	0,96%	0,30%	0,93%	1,28%	0,73%	0,08%	0,64%
GUAVIARE	0,44%	0,04%		32,15%	0,08%	11,43%			1,84%		0,06%		0,08%
HUILA	2,86%	0,09%	0,67%	4,63%	3,32%	3,45%	1,86%	5,25%	2,15%	1,98%	0,40%	0,00%	0,13%
LA GUAJIRA	6,18%	0,06%	19,62%	9,84%	1,93%	11,07%	0,80%	2,34%	0,66%	0,79%	0,98%		2,23%
MAGDALENA	6,60%	0,00%	8,87%	5,78%	1,74%	0,54%	0,10%	0,01%	0,54%	0,20%	12,85%	0,00%	1,90%
META	2,62%	0,14%	2,62%	12,02%	2,18%	2,76%	0,08%	0,03%	26,57%	0,25%	0,25%	0,00%	0,35%
NARIÑO	8,64%	0,01%	2,36%	9,80%	0,35%	0,18%	2,15%	0,30%	0,05%	6,45%	0,56%	0,03%	0,85%
NORTE DE SANTANDER	2,90%	0,64%	28,25%	2,17%	5,68%	3,01%	0,31%	0,04%	0,18%	0,19%	8,26%	0,02%	0,73%
PUTUMAYO	0,71%	0,02%	0,76%	10,90%	0,21%	1,37%	0,06%	0,91%	5,37%	2,02%	2,05%	0,04%	0,11%
QUINDÌO	10,93%	0,02%	2,12%	11,92%	5,20%	0,68%	5,32%	0,13%	0,17%	0,68%	0,71%	0,09%	0,74%
RISARALDA	9,24%	0,02%	3,37%	9,68%	2,50%	0,51%	11,99%	0,10%	0,06%	1,17%	0,15%	1,16%	0,23%
SANTANDER	6,07%	0,79%	4,96%	5,91%	7,63%	4,64%	0,66%	0,04%	1,98%	0,19%	9,31%	0,01%	0,53%
SUCRE	41,62%	0,26%	8,60%	0,64%	6,10%	0,57%	0,26%	0,01%	0,06%	0,20%	1,20%	0,02%	25,13%
TOLIMA	21,36%	0,01%	1,74%	12,37%	1,12%	0,67%	2,38%	0,95%	0,89%	1,09%	0,90%	0,29%	1,12%
VALLE DEL CAUCA	9,01%	0,05%	1,95%	11,35%	0,68%	0,68%	2,13%	0,19%	0,43%	5,02%	0,12%	0,12%	0,16%
Total	**13,54%**	**0,24%**	**7,31%**	**10,40%**	**4,40%**	**3,40%**	**1,49%**	**0,27%**	**2,28%**	**2,00%**	**2,27%**	**0,14%**	**1,51%**

Fuente: elaboración propia.

Como complemento a la construcción de la matriz origen-destino y con el objeto de apoyar la toma de decisiones gubernamentales, es relevante conocer el impacto en el transporte de carga modo carretero durante la emergencia sanitaria causada por la pandemia. Por lo anterior, se implementan dos estrategias para evaluar el impacto:

1

Un modelo estadístico que permita estimar el comportamiento de algunas variables asociadas al transporte de carga aplicando series de tiempo ARIMA;

2

Una línea de tendencia de datos antes de la COVID-19 y posterior a la reactivación o apertura gradual e inteligente durante la COVID-19 (DNP & CAOBA, 2020).

3.2.5. ANÁLISIS DESCRIPTIVO

En su inicio se realizó un análisis descriptivo del transporte carretero en Colombia. Se tomaron como entrada los datos nacionales consolidados, por municipio origen y municipio destino. A partir de lo anterior, se encontró que hubo una afectación al transporte carretero en 2016 por el paro nacional camionero y en el periodo marzo 2020 a agosto 2020, debido a la cuarentena estricta por la COVID-19. El análisis evidencia un incremento del transporte en los meses posteriores al evento —posiblemente causado por represamiento de carga—. Las variables observadas de *Viajes totales y Volumen kg* están altamente correlacionadas. (figura III.2-2).

Figura III.2-2. Análisis gráfico de la serie de tiempo para Totales (viajes, galones, kg) Colombia, 2015 (enero) - 2020 (septiembre)

Fuente: elaboración propia.

3.2.5.1. Modelo estadístico

Se adelantó un pronóstico a partir de un modelo ARIMA para estimar el impacto de la COVID-19 sobre el *total de viajes, el total de peso en kilogramos (kg), el volumen en galones (gal), el volumen en kilogramos (kg) para un municipio origen* (Buenaventura) y el volumen en kg para un municipio destino (Bogotá).
El pronóstico de impacto de la COVID-19 se practicó mediante el GAP entre los valores que se predicen con la serie ARIMA y los valores reales observados entre marzo y septiembre de 2020.

Las series para las variables de *total de viajes y volumen en kilogramos,* además de ser variables correlacionadas, fueron bastantes estables para utilizarlas con ARIMA. Sin embargo, la variable volumen en galones no tuvo buena estabilidad para generar una buena predicción, pues los intervalos de confianza en la predicción fueron dispersos.

El modelo fue efectivo para las siguientes variables:

a. Total viajes por país

b. Total en peso en kg por país

c. Total en volumen en gal por país

d. Total volumen en kg para un municipio origen (Buenaventura)

e. Total volumen en kg para un municipio destino (Bogotá)

En general, la predicción podría mejorarse posteriormente con más periodos históricos y más periodos relacionados con la nueva realidad de la COVID-19; lo anterior, puede hacerse para hallar de manera más precisa el impacto de la pandemia sobre el transporte de carga.

3.2.5.2. Línea de tendencia

Para la estimación de las líneas de tendencia de recuperación de la COVID-19 se utilizaron datos del periodo de normalidad (Pre-COVID 19) en donde se consideraron dos periodos: junio de 2017 a febrero de 2020 y otra segunda línea de tendencia considerando los últimos 12 meses Pre-COVID-19 de marzo del 2019 a febrero de 2020.
El cálculo de las líneas de tendencia se elaboró a partir de una regresión lineal simple de la variable observada en comparación con el tiempo para obtener las ecuaciones de las líneas (DNP & CAOBA, 2020).

No obstante, los esfuerzos del equipo para la construcción de las líneas de tendencia, se concluyó que es necesario contar con más datos en la nueva normalidad, de al menos 6 meses, tanto para medir y comparar la nueva línea de tendencia como para hacer estimaciones sobre el periodo en el cual se podrían alcanzar los niveles de normalidad y crecimiento que se tenían antes de la pandemia.

A partir de los dos modelos, se obtuvieron datos que sirvieron de insumo para el análisis del impacto de la COVID-19 en todo el país, las ciudades principales y los puertos de Colombia. Estos resultados están en términos del valor real, predicción, GAP —diferencia entre el valor real y la predicción— y los intervalos de confianza para los meses entre marzo y septiembre 2020 de acuerdo con las variables previamente definidas —volumen en kg, volumen en gal y cantidad de viajes totales—.

3.2.6. CONCLUSIONES Y RECOMENDACIONES

La matriz origen-destino, el desarrollo de la herramienta para la visualización de los datos en Power BI, el análisis del efecto en el transporte carretero de la pandemia por la COVID-19 mediante serie de tiempo ARIMA y el análisis de las líneas de tendencia, permiten a los tomadores de decisiones apoyar sus propuestas basadas en el análisis de datos como un elemento complementario a otras estrategias, ayudando no solo a entender el impacto de la COVID-19, sino también a anticipar acciones para enfrentra sus efectos (DNP & CAOBA, 2020).

El modelo actual tiene oportunidades de mejora dado que hubo poca disponibilidad de datos para hacer un pronóstico más ajustado. Por ello, el modelo actual presenta limitaciones para evidenciar que el cambio en el comportamiento de la carga a lo largo del país durante el período de análisis se deba a la COVID-19. Como se identifican tres escenarios diferentes, a saber:

1 Antes de la COVID-19,

2 La COVID-19 con cuarentena permanente y

3 La COVID-19 con la "nueva normalidad", para los escenarios 2 y 3 no hay datos suficientemente representativos del nuevo comportamiento ni para determinar lo que se sugeriría para el transporte de carga por carretera.

Por último, con la finalidad de ajustar la predicción del modelo ARIMA, se recomienda evaluar la correlación de los datos con variables macroeconómicas y sociales, para evaluar y diseñar escenarios de impacto y de recuperación de la COVID-19.

CAPÍTULO 3.3

PRONÓSTICO DE LA DEMANDA Y EL COSTO ASOCIADO AL SERVICIO DE CUIDADOR PERMANENTE EN SALUD[13]

3.3.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

Todas las personas sin excepción requieren de cuidado en todas las etapas de la vida. Este es entendido como una condición que es imprescindible para la sostenibilidad de la vida humana y también para dinamizar los procesos de bienestar personal y colectivo. Por lo anterior, para que el cuidado sea garantizado, debe ser comprendido como una actividad que requiere de una agenda política integrada por acciones del Estado, de las familias, de las empresas privadas y la comunidad. Aunque todas las personas requieren de cuidado, este puede ser más demandante para grupos poblacionales como los niños y niñas menores de cinco años, los adultos mayores, y las personas con algún tipo de dependencia.

En el contexto de la pandemia se presenta un aumento de personas que requieren de labores de cuidado por presentar patologías vulnerables a la COVID-19. Esa situación genera un reto para el Sistema de Protección Social de Colombia, dado que no se conoce con precisión el costo del servicio de cuidado del sector salud así como estimación de su demanda, motivo por el cual se hace complejo direccionar y focalizar recursos públicos de manera adecuada (DNP & CAOBA, 2020).

13. Integrantes del equipo: Andrés Moreno Barbosa, María Camila Martínez, José Francisco Molano, Jairo Tirado, Nicolás Agudelo.

3.3.2. OBJETIVOS DEL PROYECTO

Objetivo general

Describir y pronosticar la demanda y el costo para el componente de salud de las atenciones relacionadas con el cuidado para la población cuyo diagnóstico requiere de cuidado permanente.

Objetivos específicos

- Realizar el reporte técnico con los diagnósticos relacionados para considerar en el estudio, específicamente los que se relacionan con labores de cuidado en el componente de salud.
- Caracterizar las poblaciones vulnerables impactadas por la COVID-19 que requieren labores de cuidado dada las patologías y comorbilidades existentes.
- Describir y comparar los costos y la demanda de los servicios de cuidado diferenciados por régimen —subsidiado y contributivo—.
- Describir los costos y la demanda de las atenciones relacionadas con el cuidado por regiones a nivel de municipio.
- Elaborar un modelo de pronóstico de la demanda y el costo del cuidador para poblaciones caracterizadas.

3.3.3. DATOS UTILIZADOS

El equipo experto identificó y gestionó para el proyecto las fuentes de datos presentadas en la tabla III.3-1.

Tabla III.3-1. Fuentes de datos utilizadas

CATEGORÍA	FUENTE	DESCRIPCIÓN
Prestación de servicios de salud	Registro Individual de Prestación de Servicios en Salud (RIPS)	Se registra la información de identificación personal, tipo de evento, diagnósticos principales y secundarios, municipio, fecha de atención, edad y costos de procedimientos.
Plan de Beneficios en Salud	Datos de Suficiencia	Cubo de datos del Ministerio de Salud con el registro de procedimientos, tecnologías y medicamentos incluidos en el plan de beneficios en salud con cargo a la unidad de pago por captación. La base de datos resuelve consultas acerca del total de personas únicas que tuvieron procedimientos, tecnologías o medicamentos autorizados para una determinada patología.
Datos de recobros	Base de datos de recobros ADRES	Contiene la información de registros individuales de reconocimiento y pago de los servicios médicos y/o los medicamentos no incluidos en el Plan Obligatorio de Salud prestados a los afiliados o beneficiarios, de acuerdo con lo establecido en la normativa vigente. La base de datos contiene la información referente a la entidad prestadora de salud, la información personal del paciente (sin la identificación personal), la fecha de prestación del servicio, descripción del servicio prestado e información de los costos correspondientes a los mismos.

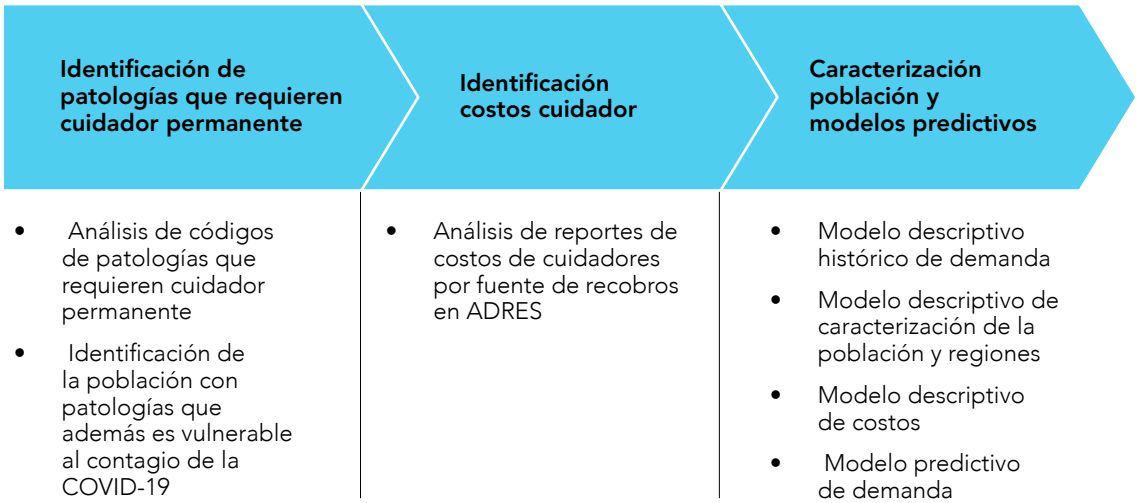
Fuente: elaboración propia.

A su vez, el equipo técnico de este proyecto construyó la base de datos de patologías que requieren cuidador permanente, a través de la identificación y caracterización de patologías.

3.3.4. MODELOS Y RESULTADOS PRINCIPALES

El análisis del problema de política pública conlleva a la necesidad de establecer una estrategia para caracteriza a la población que requiere servicios de cuidado y los costos asociados a estos servicios.
Esta estrategia sigue tres pasos principales presentados en la figura III.3-1.

Figura III.3-1. Estrategia de solución del proyecto

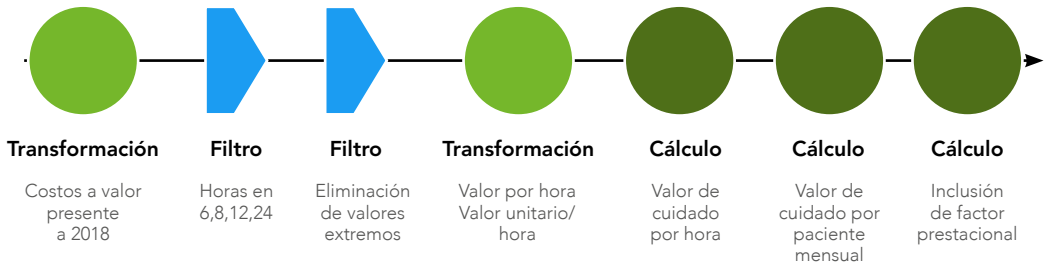


Fuente: elaboración propia.

En primer lugar, se clasificaron y caracterizaron patologías que requieren cuidador permanente de acuerdo con criterios médicos; de las inicialmente clasificadas, se identificaron sus respectivos códigos CIE[14] 10 y se seleccionaron 1.884 de ellas[15]. En segundo lugar, se identificaron de costos de cuidado, siguiendo el flujo presentado en la figura 3-2. Así se analizaron los recobros autorizados relacionados con los cuidadores para determinar el costo por hora de un cuidador y extrapolar ese costo para hallar los que tendrían que asumirse para atender la demanda observada.

El proceso aplicado para la de definición de costos se define en el flujo presentado a continuación (figura III.3-2).

Figura III.3-2. Flujo de tareas para la estimación del costo de cuidado por paciente y por mes



Fuente: elaboración propia.

14. Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud.
15. Estas patologías incluyen enfermedades huérfanas, tumores, otras, transmisibles, emergentes, reemergentes y desatendidas, cardiovasculares, cáncer, psiquiatría y neurología, diabetes, sanguíneas y respiratorias.

La estimación permitió concluir que el valor unitario por hora corresponde a $6.500. A partir de ese resultado se estimó el valor mensual por paciente obtenido de multiplicar 8 horas por 25 días, e incluir el factor prestacional. El cálculo anterior permitió arrojar un valor de $1.800.000 como el precio de cuidador. Finalmente, se implementaron los modelos para la caracterizar la población y modelos predictivos de la demanda.

3.3.4.1. Modelos descriptivos de demanda y costos

A partir de los modelos descriptivos generados usando la variable de conteo escogida con el equipo del DNP se pudo caracterizar el estado de las patologías que requieren cuidado especial.
El conteo se entendió como usuarios únicos que acudieron a un servicio de cuidado en una granularidad de tiempo (año o año-mes).

Los modelos descriptivos ofrecen al usuario final los siguientes resultados:

1. Entender con mayor facilidad la evolución de la trazabilidad de los conteos de diagnósticos de acuerdo con la vulnerabilidad a la COVID-19, régimen, tipo costo, tipo enfermedad y grupo etario,
2. Comprender los cambios de estructura poblacional por entidad territorial y
3. Analizar las diferencias entre entidades territoriales respecto a su acumulación de diagnósticos de cuidado por medio de una métrica de normalización como el Z-score.

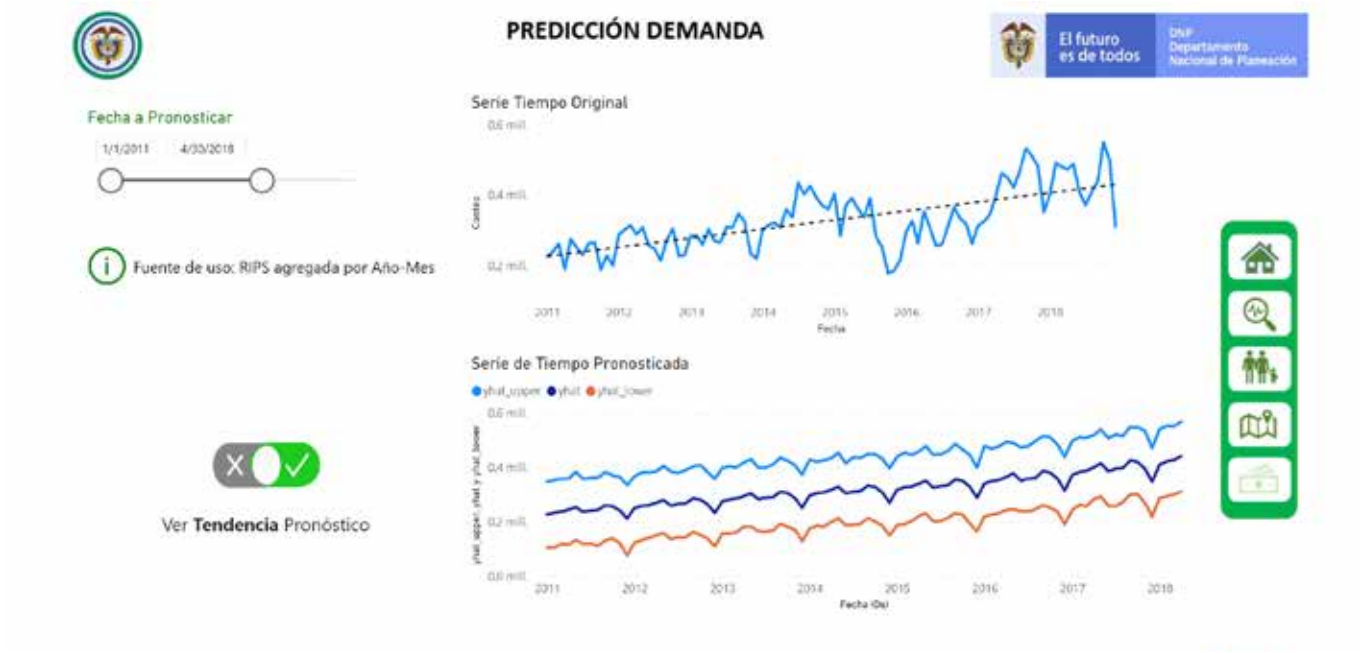
El análisis de estructura poblacional que se desprende del modelo predictivo es clave para definir planes de prevención y gestión de patologías relacionadas con cuidado permanente en los territorios, lo anterior dado que permite la caracterización de la población por sexo y grupo etario.

3.3.4.2. Modelos predictivos

Por otro lado, para el modelo predictivo de la demanda se propuso un modelo de predicción de series de tiempo mensuales de personas únicas que han accedido a servicios de salud. (DNP & CAOBA, 2020) (figura 3-3). A través del modelo se logró estimar la predicción de la demanda de servicios de salud, la cual tiene una gran varianza en la serie de origen, por lo que la predicción que se ajusta como resultado ofrece un intervalo de confianza con una gran amplitud —aproximadamente 150.000 personas en ambas direcciones en todo el país—.

Al utilizar el modelo de demanda estimado por la tendencia, se espera que para los próximos años la población base que necesita cuidador permanente en el país se incremente a 608.024 personas a diciembre del año 2022. Y al extrapolar costos de cuidador permanente para estas personas, se estima un costo de $109.443 millones en el mes de diciembre de 2022. (figura III.3-3).

Figura III.3-3. Serie de tiempo de personas únicas por mes con patologías que requieren servicios de cuidado que han tomado servicios de salud durante el periodo 2011-2018



Fuente: tomado de informe final (DNP & CAOBA, 2020).

Tabla III.3-2. Predicción de tendencia de la serie de tiempo y costo asociado para meses diciembre 2020, 2021 y 2022

MES	PREDICCIÓN TENDENCIA	COSTOS EN MILLONES DE PESOS($)
Diciembre de 2020	528.693	951.647
Diciembre de 2021	568.359	1.023.046
Diciembre de 2022	608.024	1.094.443

Fuente: elaboración propia.

3.3.5. RECOMENDACIONES

Se identifica la necesidad de mejorar la precisión de la lista de patologías que se deben considerar para la agregación de la base de datos de RIPS, y para estimaciones posteriores se recomienda la actualización de los valores Z-Score de los municipios, que actualmente se calcula con Censo DANE, incluyendo proyecciones de la población desagregada por municipios para los años establecidos en el análisis.

En relación con la estimación de los costos de cuidado, es recomendable hacer un análisis de manera diferenciada por diagnóstico dado que algunos diagnósticos presentan una varianza del costo más elevada y su promedio es mayor.

CAPÍTULO 3.4

DINÁMICA DEL EMPLEO FORMAL Y EMPRESAS BAJO LA CONTINGENCIA DE LA COVID-19[16]

3.4.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

En respuesta a la emergencia económica y social derivada de la crisis de la COVID-19, el Gobierno colombiano adoptó medidas para proteger los puestos de trabajo y el tejido empresarial. Entre las determinaciones mencionadas se encuentra el Decreto 558 de 2020, que tuvo como objetivo brindar, a través del Sistema General de Pensiones, una mayor liquidez a los empleadores y trabajadores dependientes e independientes (DNP & CAOBA, 2020). En el artículo 3 de este Decreto se establece un pago parcial del aporte al Sistema General de Pensiones para los períodos de abril y mayo de 2020. En consecuencia, los empleadores de los sectores público y privado al igual que los trabajadores independientes al optar por este alivio pagarían como aporte solo un 3% de cotización al Sistema General de Pensiones. Esta cotización fue pagada en un 75% por el empleador y un 25% por el trabajador, por su parte en caso de trabajadores independientes lo debían hacer por el 100%.

Con el fin de monitorear la efectividad de la medida descrita, la Dirección de Innovación y Desarrollo Empresarial (DIDE) propuso desarrollar una herramienta para dar seguimiento a los cambios asociados a este decreto por medio de la observación simplificada y resumida de los datos que componen el Sistema de Seguridad Social a través de la Planilla Integrada de Liquidación de Aportes (PILA) y el Registro Único Empresarial y Social (RUES).

16. Integrantes del equipo: María Piedad Bayter Horta, Lorena Andrea López Barrera, Jaime Andrés Pavlich-Mariscal, Andrés Rodríguez, Juan David Rodríguez.

3.4.2. OBJETIVOS DEL PROYECTO

Objetivo general

Caracterizar la dinámica del empleo formal y de las empresas de Colombia, a raíz de la contingencia de la COVID-19 y el Decreto 558 de 2020 del Gobierno nacional.

Objetivos específicos

- Cuantificar la variación del empleo en las empresas, de acuerdo con su tamaño, sector económico y ubicación.
- Cuantificar la variación en las novedades de vacaciones y licencias de los empleados de las empresas.
- Cuantificar las empresas beneficiarias de las medidas del Decreto 558, de acuerdo con su tamaño, sector económico y ubicación.
- Estimar el ahorro a las empresas beneficiarias de las medidas del Decreto 558.

3.4.2.1. Datos utilizados

La tabla III.4-1 describe los datos utilizados por este proyecto, que básicamente se nutrió de la base PILA (Planilla Integrada de Liquidación de Aportes) y la base RUES (Registro Único Empresarial y Social), más de los descriptores geográficos que permitieron las visualizaciones en mapas y los códigos de clasificación de actividad económica, que fueron útiles para agrupar los datos según distintas ramas de actividad.

Tabla III.4-1. Fuentes de datos utilizadas

CATEGORÍA	FUENTE	DESCRIPCIÓN
Aportantes y cotizantes	Base de datos Planilla Integrada de Liquidación de Aportes (PILA)	Aproximadamente 830.000 aportantes (empleadores) y 13.500.00 cotizantes (empleados)
Empresas formales	Base de datos de Registro Único Empresarial y Social	Aproximadamente 1.600.000 registros de empresas
Información geográfica	Base de datos de la División Político-Administrativa de Colombia (DIVIPOLA)	Coordenadas geográficas de departamentos/ municipios • Formato: XLS • Acceso: https://geoportal.dane.gov.co/ • 7978 registros de centros poblados en Colombia
Actividad económica	Base de datos con códigos CIIU (Clasificación Industrial Internacional Uniforme)	Describe actividades económicas y sus códigos CIIU

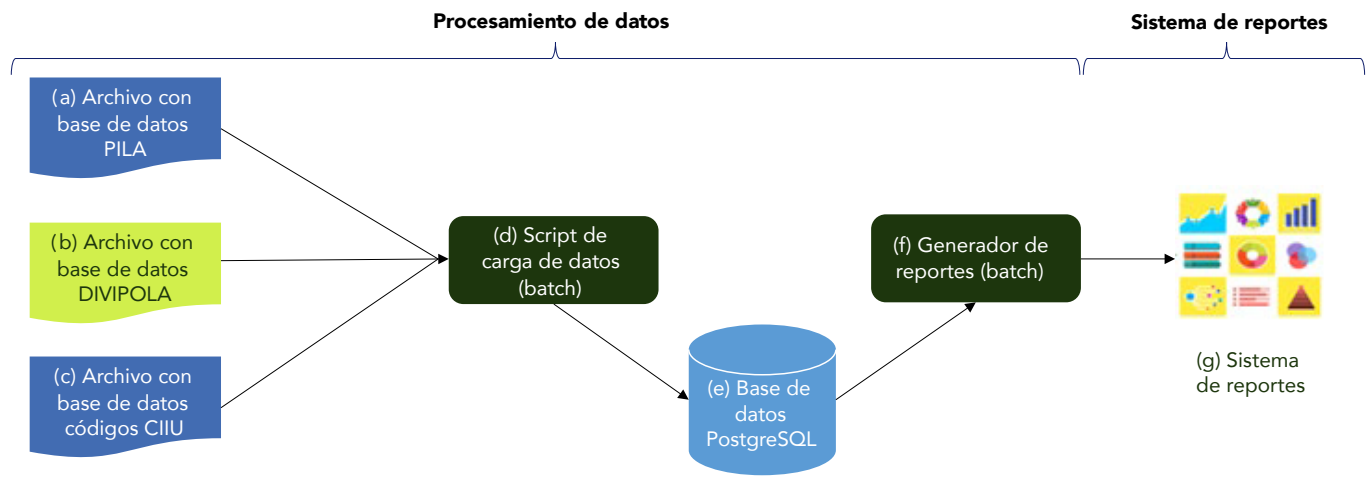
Fuente: elaboración propia.

3.4.3. SOLUCIÓN PROPUESTA

La figura III.4-1 detalla los componentes de la solución propuesta que tiene dos secciones: *procesamiento de datos (a-f) y sistema de reportes (g).* El procesamiento de datos solo se ejecuta una vez por parte del equipo de investigación y desarrollo, mientras que el sistema de reportes es el que los usuarios finales pueden utilizar repetidamente y de manera directa.

Un *script* escrito en PostgreSQL *(d)* carga los archivos *(a), (b) y (c),* descritos en la figura III.4-1, en una base de datos PostgreSQL *(e).* Esta última será usada por el proceso *batch (f),* escrito en Python, para limpiar los datos y generar los reportes. El sistema de reportes *(g)* está implementado en HTML, CSS y Javascript, basado principalmente en la librería Plotly.

Figura III.4-1. Componentes de la solución

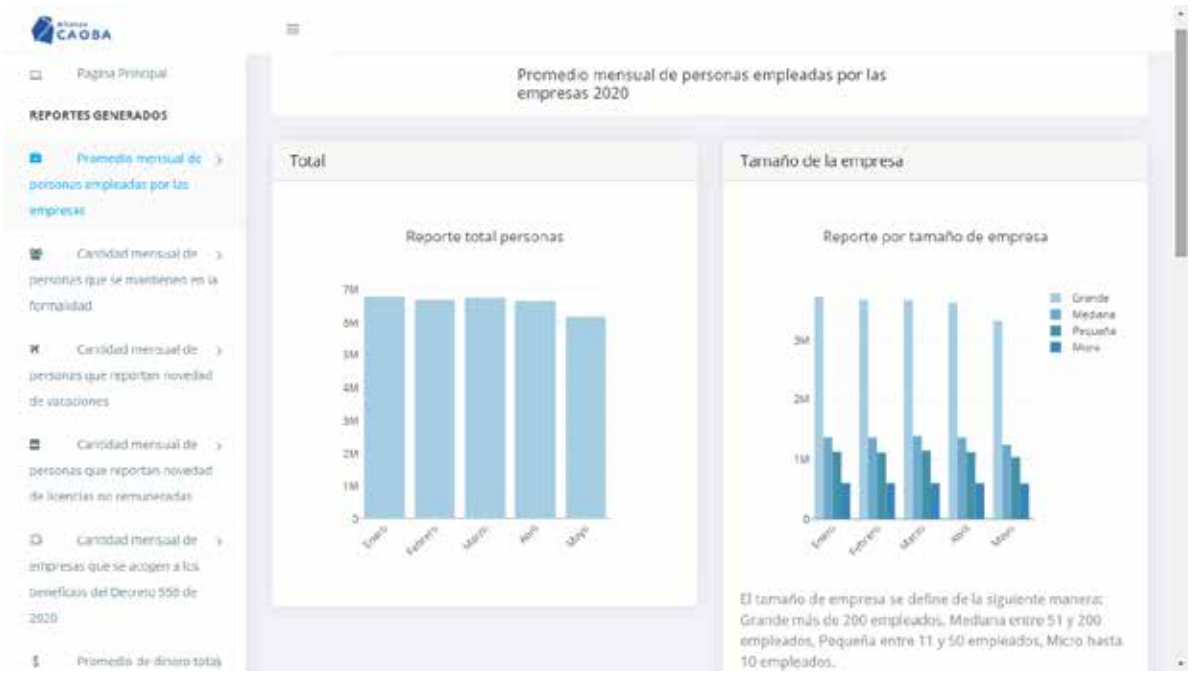


Fuente: tomado de informe final (DNP & CAOBA, 2020).

3.4.4. RESULTADOS

El sistema desarrollado es capaz de generar seis tipos de reportes (figura III.4-2) a través de una aplicación con la que pueden interactuar los usuarios finales del sistema. Entre estos reportes se destaca el que muestra promedios mensuales de trabajadores formales, el cual incluye un filtro para ajustar la cantidad de determinantes de ese estatus de formalidad (salud, pensión y ARL).

Figura III.4-2. Tablero de visualización



Fuente: tomado de informe final (DNP & CAOBA, 2020).

Por otro lado, cada reporte utiliza distintos tipos de visualizaciones, que incluyen gráficos de barras con diferentes clases de aperturas —por tamaño de empresa, sector de actividad y departamento de ubicación— y mapas de calor por meses y departamentos; también genera gráficos por tamaño, sector de empresa y departamento que tienen opciones para filtrar información. Por ejemplo, en la visualización de sectores económicos el sistema provee una interfaz para seleccionar arbitrariamente los sectores por desplegar en el gráfico con una desagregación que permite analizar actividades económicas definidas con precisión. De manera similar, el gráfico de departamentos permite seleccionar los departamentos y compararlos en el tiempo, al igual que el mapa de calor provee facilidades de *zoom y panning*, un filtro por mes y la posibilidad de excluir explícitamente departamentos para facilitar la visualización en presencia de *outliers*.

3.4.5. RECOMENDACIONES

El sistema desarrollado tiene el potencial de ser ampliado para futuros proyectos. En particular, este sistema puede adaptarse fácilmente para incorporar información actualizable de manera automática y en tiempo casi real —a medida que se generen los reportes en los datos administrativos que proveen insumos al sistema—. Este tipo de automatización podría otorgar a los decisores de política de información de seguimiento detallada por tipo de empresas, sector y su ubicación geográfica, que puede resultar de gran utilidad para la toma de decisiones acerca de políticas de desarrollo productivo, ya sea en un contexto de crisis económica como el impuesto por la pandemia de la COVID-19, como para períodos de reactivación y normalidad.

CAPÍTULO 3.5

ANALÍTICA DE DATOS PARA ESTIMAR EL RIESGO DE **DESNUTRICIÓN DE NIÑOS Y NIÑAS EN COLOMBIA,** EN EL MARCO DE LA EMERGENCIA POR LA COVID-2019[17]

3.5.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

La desnutrición es un problema grave que aqueja a una porción importante de los niños en Colombia. Se caracteriza por un deterioro de la composición corporal y una alteración sistemática de las funciones orgánicas y psicosociales que pueden poner en riesgo su vida. De acuerdo con los datos del año 2015 (ENSIN, 2015), la prevalencia de la desnutrición crónica en el país en menores de 5 años era del 10,8% y la de la desnutrición aguda del 1,6%.

La llegada de la pandemia, y con ella la de las medidas de distanciamiento social impuestas para contenerla, amenazan con agravar la situación existente. De acuerdo con UNICEF, la desnutrición infantil tiene significativas interacciones con la crisis por la COVID-19 por al menos tres razones.

1 Por la posible reducción o limitación en la provisión de servicios de asistencia alimentaria.

2 Porque la malnutrición puede agravar el riesgo de padecer COVID-19.

3 Porque los niños dependen de sus padres y cuidadores para su adecuada nutrición y en caso de que ellos se enfermen, tengan que permanecer confinados o debido a cambios drásticos en las dinámicas familiares durante la pandemia, muchos niños en riesgo de desnutrición pueden ver agravada esa condición.

17. Integrantes del equipo: Carolina Suárez, Manuel Reina, Catalina González, Carlos Díaz, Julio Carreño, Mónica Vargas, Daniel Nieto.

Además, durante la pandemia el riesgo de desnutrición infantil puede aumentar debido a la caída en los ingresos de las familias más vulnerables y a interrupciones en los servicios de salud y de protección social a los niños en estas familias (Headey *et al.*, 2020).

En este contexto, el presente proyecto intenta brindar al país una herramienta para identificar a las niñas y los niños en riesgo de sufrir desnutrición aguda. Esta información puede resultar valiosa en el momento de definir políticas tanto preventivas como paliativas del fenómeno de interés.

3.5.2. OBJETIVOS DEL PROYECTO

Objetivo general

Identificar el riesgo de desnutrición de niñas y niños en primera infancia y el impacto de la COVID-19 en esta situación, mediante la construcción de un modelo predictivo basado en datos relevantes disponibles que apoye la toma de decisiones.

Objetivos específicos

- Estimar para cada niño y niña el riesgo de desnutrición mediante un modelo de analítica de datos.
- Estimar el impacto de la COVID-19 en el riesgo de desnutrición.
- Identificar brechas de política pública en la atención de niños y niñas con riesgo de desnutrición.

3.5.3. DATOS UTILIZADOS

Los datos utilizados provienen de diferentes fuentes, entre las que se destacan:

Tabla III.5-1. Fuentes de datos utilizadas

CATEGORÍA	FUENTE	DESCRIPCIÓN
Atención y prevención de la desnutrición	Base cuéntame del Instituto Colombiano de Bienestar Familiar	Sistema de información para apoyar la gestión y recolección de información de los servicios que ofrece la Dirección de Primera Infancia
Condiciones socioeconómicas de los hogares y las personas	Sisbén IV	
Información geográfica	Base de datos de la División Político-Administrativa de Colombia (DIVIPOLA)	Coordenadas geográficas de departamentos/municipios 7978 registros de centros poblados en Colombia

Fuente: elaboración propia.

Adicional a las fuentes de información empleadas para el análisis de este proyecto, se construyó una variable objetivo de estado-peso-talla, a partir de la cual se definieron tres categorías que distinguen grados de desnutrición infantil:

- Desnutrición aguda severa
- Desnutrición aguda moderada
- Riesgo desnutrición aguda

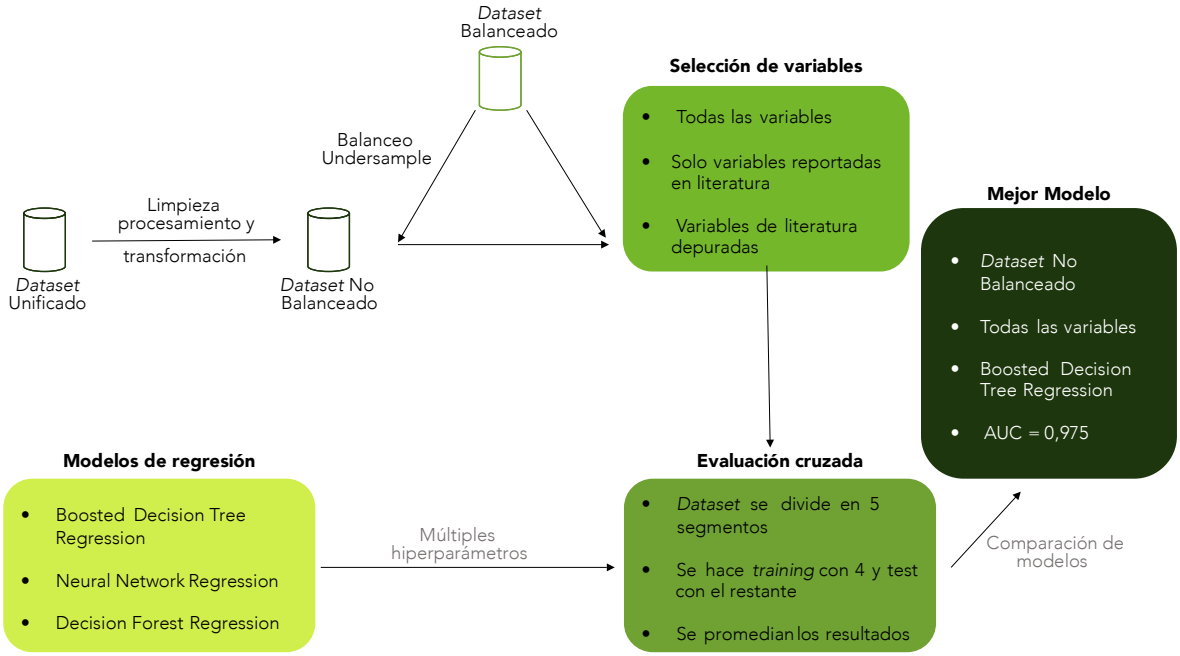
Otras variables de interés para el proyecto fueron las siguientes: tiempo de lactancia materna, grupo étnico, discapacidad, medidas antropométricas varias, características de la vivienda, del jefe del hogar, acceso a servicios públicos y agua potable, ingresos y gastos del hogar, etc. También se incluyó en el análisis el nivel de afectación por la COVID-19 por municipios, a partir de datos del MinSalud.

3.5.4. SOLUCIÓN PROPUESTA

Luego de los consensos alcanzados en el *workshop* inicial de Manos en la Data y de las reuniones posteriores con representantes del ICBF, el MinSalud, la Fundación Éxito, y el Programa Mundial de Alimentos, se procedió a una revisión exhaustiva del estado de la literatura que analiza los determinantes de la desnutrición infantil y propone modelos de análisis de ese fenómeno. La revisión alimentó las etapas de selección de variables y de posibles modelos por evaluar, dentro del proceso que se describe en la figura III.5-1.

En paralelo, se fueron consolidando las bases de datos y se inició el análisis que consistió en la limpieza inicial y consolidación posterior de las bases de datos y en la inclusión de las variables seleccionadas en los modelos de regresión propuestos (Boosted Decision Tree Regression, Neural Network Regression, Decision Forest Regression), los cuales pasaron por un proceso de validación cruzada para, finalmente, elegir el que resultó ser el mejor modelo: Boosted Decision Tree Regression, AUC = 0,975.

Figura III.5-1. Proceso del modelamiento del proyecto



Fuente: tomado de informe final (DNP & CAOBA, 2020).

Luego se procedió de manera similar, pero incorporando información sobre afectación por la COVID-19 por municipio. De esta manera se generaron dos *datasets* enriquecidos: uno que contiene la predicción de grado de desnutrición sin tener en cuenta la afectación por la COVID-19 y otro con una predicción que sí la tiene en cuenta. Después se procedió a elaborar la herramienta de visualización, utilizando Power BI, que permite llevar a cabo consultas a través de mapas y también posibilita consultas sobre la base de datos enriquecida (Sisbén IV + Cuéntame + predicción desnutrición) de modo que sirva para identificar las brechas de política pública en la atención de niños y niñas con riesgo de desnutrición.

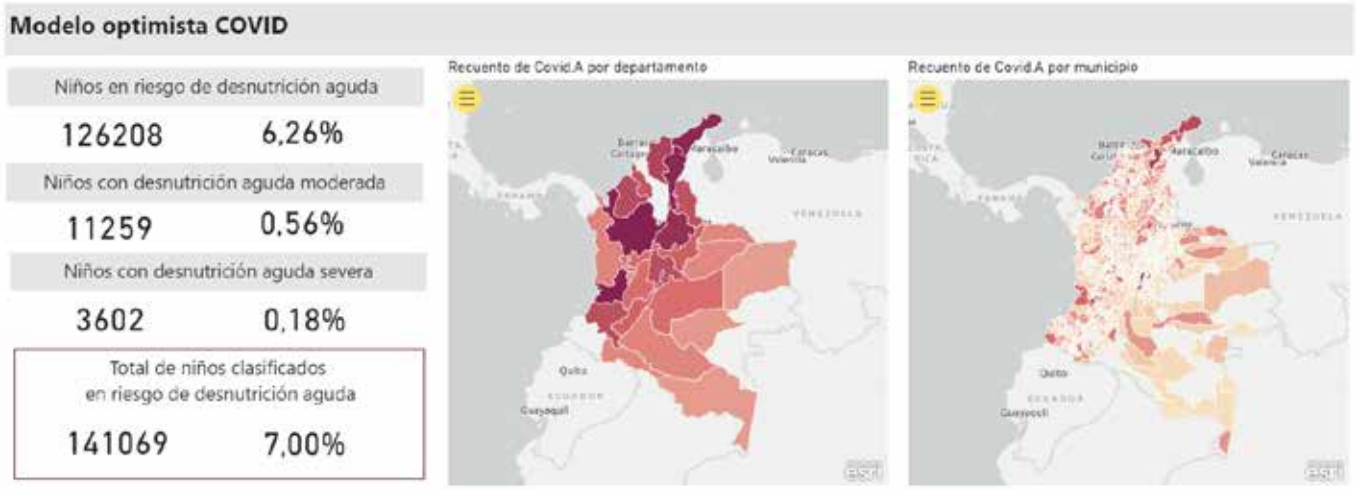
3.5.5. RESULTADOS

Como resultado de este proyecto se estimó un puntaje de riesgo de sufrir desnutrición aguda para un total de 2 millones de niñas y niños en Colombia. La herramienta desarrollada contribuye tanto a la focalización de políticas de mitigación del riesgo de desnutrición por de individuos y por territorios.

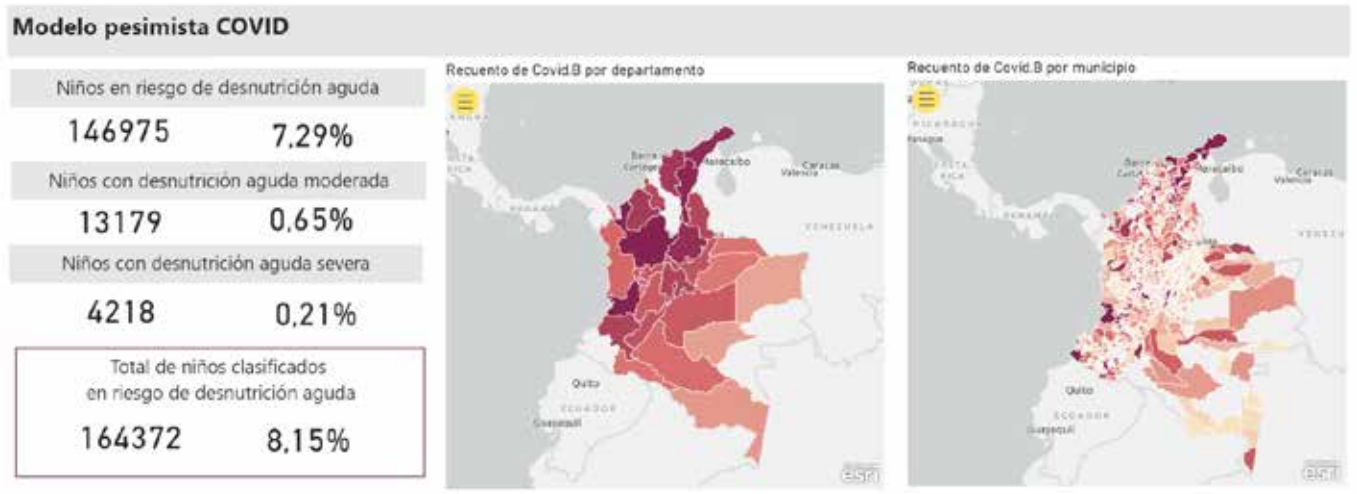
El modelo estimó que cerca de 107.000 niñas y niños se encontrarían en situación de desnutrición aguda. La estimación corresponde a 30.000 niñas y niños adicionales a los identificados anteriormente en situación de desnutrición aguda por el ICBF. Al incluir la afectación municipal por la COVID-19, se proyectaron dos escenarios donde la cifra de niñas y niños en desnutrición ascendería a 141.000 (escenario optimista, panel A figura III.5-2) y 164.000 (escenario pesimista panel B figura III.5-2).

Figura III.5-2. Mapas con indicadores de riesgo de desnutrición infantil, en dos escenarios de impactos de la pandemia

Panel A. Modelo optimista de la COVID-19



Panel B. Modelo pesimista de la COVID-19



Fuente: tomado de informe final (DNP & CAOBA, 2020).

3.5.6. RECOMENDACIONES

Del presente proyecto se desprenden las siguientes recomendaciones generales para próximos pasos:

1 Utilizar los resultados del análisis para vinculación de niñas y niños en los programas sociales del país.

2 Complementar el análisis con modelos para la desnutrición crónica y global.

3 Explorar metodologías complementarias para aislar los efectos individuales de los determinantes de la desnutrición.

4 Complementar la identificación de brechas de política pública en la atención de niños y niñas con riesgo de desnutrición, por ejemplo, el acceso de transferencias monetarias para las familias.

CAPÍTULO 3.6

EVOLUCIÓN DE LOS INDICADORES DE SEGURIDAD CIUDADANA EN CONSECUENCIA DE LAS MEDIDAS SANITARIAS[18]

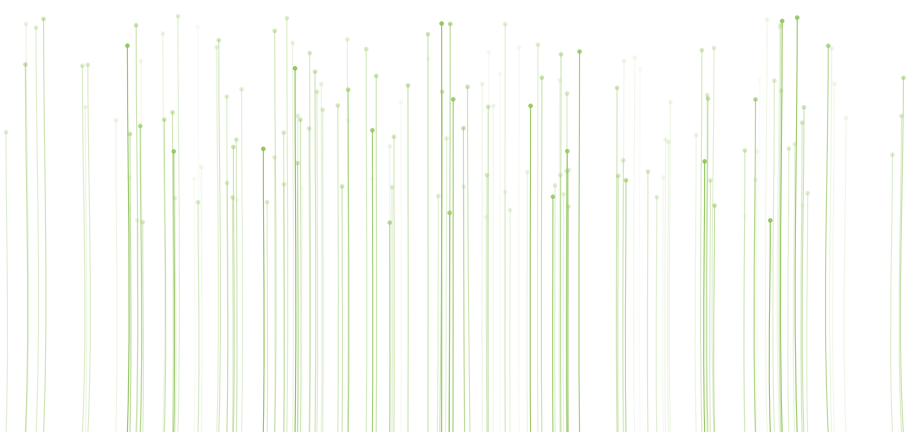
3.6.1. PROBLEMÁTICA DE POLÍTICA PÚBLICA QUE MOTIVÓ EL PROYECTO

Los eventos delictivos tienen dinámicas que difieren de acuerdo con el tipo de delito y la población afectada. También las condiciones extraordinarias impuestas por la crisis sanitaria de la COVID-19 y las medidas de distanciamiento social que rigieron para contener el avance de la pandemia afectaron esas dinámicas de los eventos delictivos.

La gran riqueza que ofrecen hoy los datos georreferenciados sobre delitos permite desarrollar análisis que ayudan a brindar luces sobre las complejas dinámicas de los eventos delictivos. En el contexto de Manos en la Data - Colombia, y dentro de la contingencia provocada por la pandemia de la COVID-19, el proyecto se planteó utilizar ese caudal de datos para analizar las dinámicas delictivas en Bogotá y Medellín. El alcance geográfico del proyecto abarcaba una ciudad más y otros tipos de análisis para el caso de Bogotá, pero debido a la corta ventana de tiempo para desarrollar el proyecto y a las gestiones necesarias para acceder a datos de ciudades adicionales, se optó por desarrollar el prototipo de ciencia de datos con los datos que estuvieron disponibles a las pocas semanas de iniciado el proceso de Manos en la Data. Esa determinación significó que el análisis espacial detallado solo fuera realizado para el caso de Medellín. Sin embargo, en caso de contar con información para otras ciudades, el prototipo puede adaptarse y aplicarse de manera poco costosa.

Los eventos delictivos en los que se enfocó el proyecto fueron hurto a personas, homicidio, violencia intrafamiliar y lesiones personales.

18. Integrantes del equipo: Henry Laniado, Andrés Pérez-Coronado, Mateo Bonnett, Carolina Matamoros, Juan David Gélvez Ramírez.



3.6.2. PREGUNTAS POR RESOLVER CON EL PROYECTO

El proyecto se planteó construir un modelo analítico integrado que permitiera entender la evolución de hechos delictivos, sus puntos de mayor concentración, y cómo estos puntos cambian en el marco de la emergencia sanitaria por la COVID-19. Para este fin, se formularon las siguientes preguntas:

- ¿Cuáles son los patrones básicos del comportamiento delictivo y sus tendencias para distintas zonas dentro de los centros urbanos analizados antes y durante la emergencia sanitaria?

- ¿Cuál es la relación entre el número de casos de la COVID-19, las medidas de aislamiento obligatorio y los índices de seguridad históricos y estimados?

La respuesta a estas preguntas ayudaría a mejorar la focalización, mediante el uso de un modelo analítico espacio temporal de la atención preventiva por los eventos delictivos, a escala territorial, en el marco de la emergencia sanitaria de la COVID-19.

3.6.3. DATOS UTILIZADOS

Tabla III.6-1. Fuentes de datos utilizadas

CATEGORÍA	DESCRIPCIÓN DE DATOS
Delitos priorizados	Hurto a personas, homicidios, lesiones personales, violencia intrafamiliar
	Datos georreferenciados de Medellín
	Datos por evento o caso delincuencial (Bogotá y Medellín)
Información pandemia	Contagios (Bogotá) por localidades
	Proyección de contagios y UCI
	Atención de la pandemia (datos por UPZ solo para Bogotá)
Llamadas de emergencia	Información por jurisdicción territorial (Bogotá)
	Información por caso rexportado mediante llamada
	Descripción de la llamada
Línea Púrpura	Información por jurisdicción territorial (Bogotá)
	Información por caso / llamada (por localidad UPZ)
	Descripción de la llamada (limitado por sensibilidad de la información)

Fuente: elaboración propia.

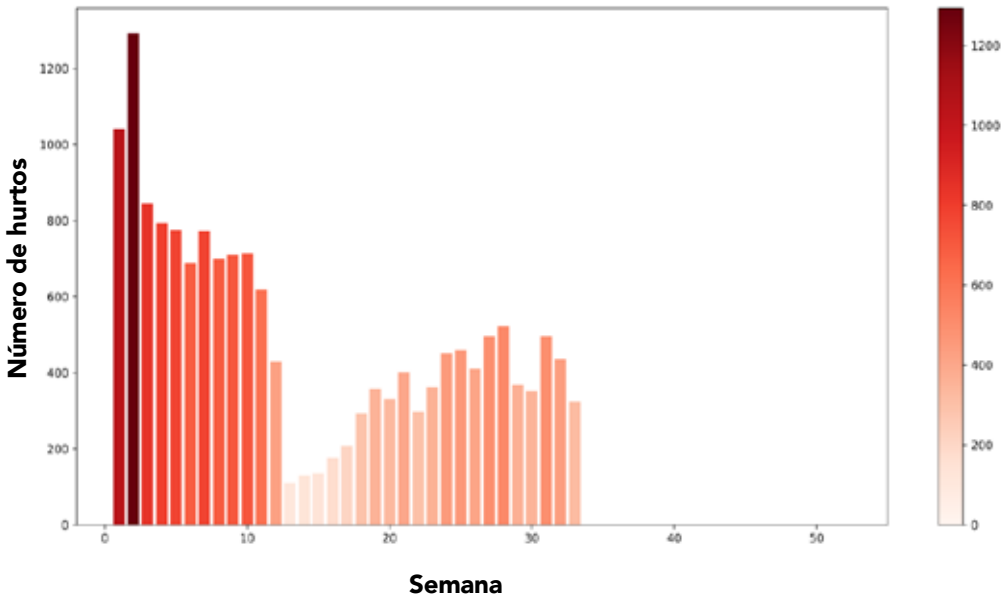
3.6.4. SOLUCIÓN PROPUESTA

3.6.4.1. Análisis descriptivo

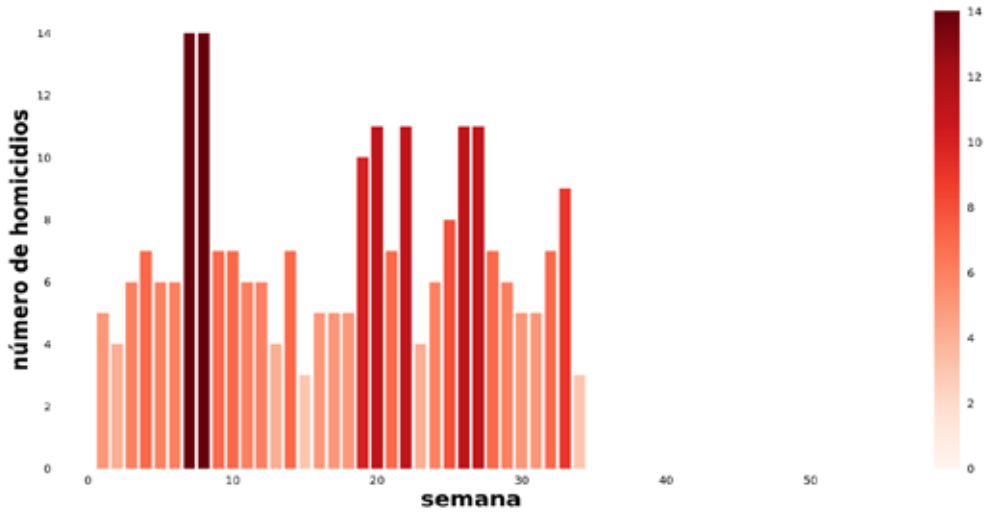
Para entender mejor las bases de datos y el comportamiento de los indicadores de seguridad en el marco de la emergencia sanitaria, se adelantó un proceso detallado de análisis descriptivo de los datos. En particular, y dada la riqueza de la información disponible, este análisis se enfocó en el caso de Medellín. Se construyeron visualizaciones y estadísticas descriptivas que permitieron observar la evolución temporal de los delitos priorizados para el análisis, al igual que como ciertos patrones de las dinámicas delictivas de acuerdo con los días de la semana y días dentro de cada mes. A su vez, este análisis dio las primeras evidencias de la relación entre los eventos delictivos y las medidas de aislamiento obligatoria impuestas a consecuencia de la pandemia de la COVID-19. La figura III.6-1 muestra un ejemplo de estas visualizaciones, comparando la distribución semanal de los hurtos en Medellín durante 2019 y las primeras semanas del año 2020.

Figura III.6-1. Distribuciones semanales del número de casos de hurto en la ciudad de Medellín, durante 2019 y primera mitad de 2020

Panel A. 2019



Panel B. 2020



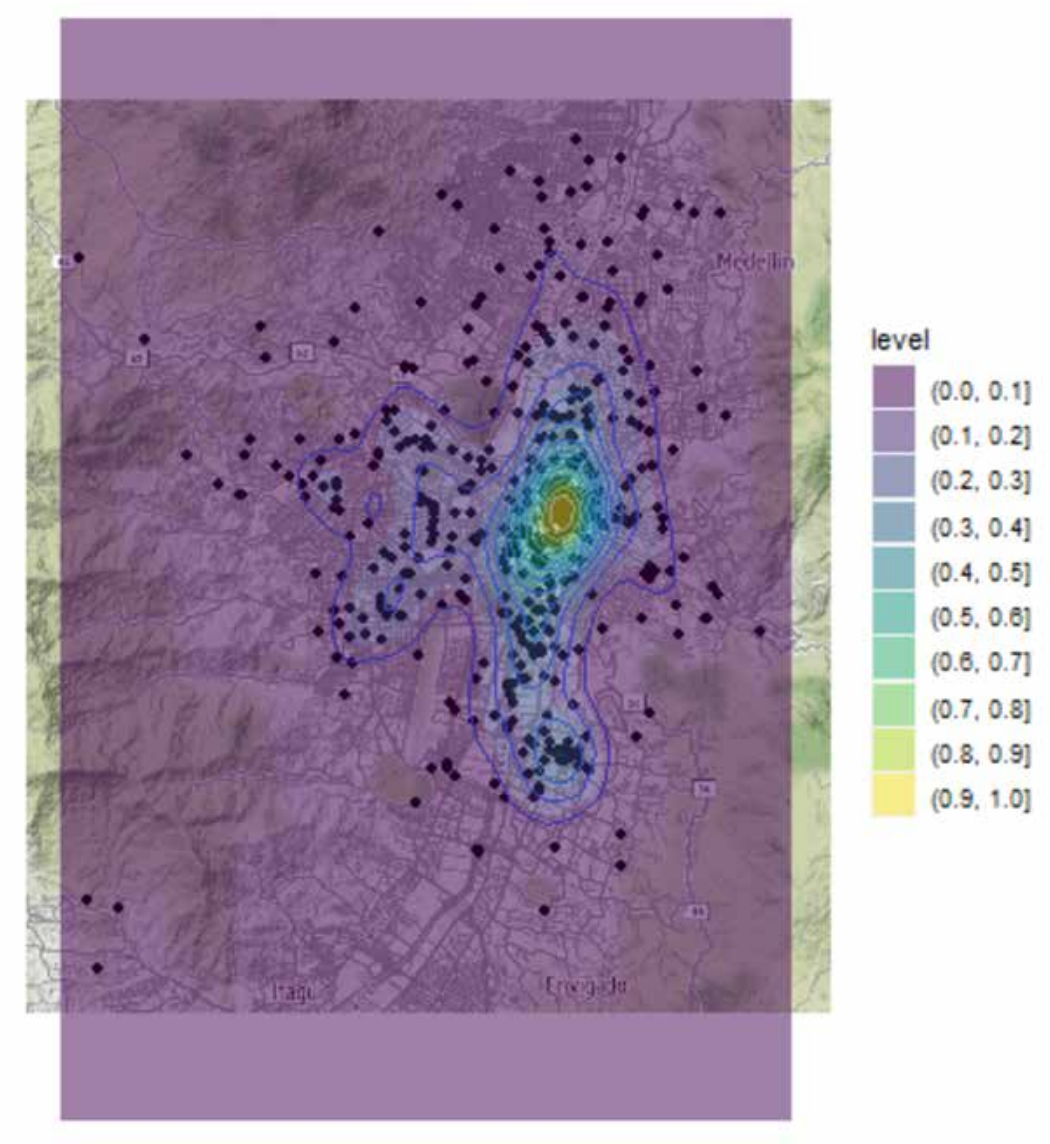
Fuente: tomado de informe final (DNP & CAOBA, 2020).

3.6.4.2. Análisis espacial

El análisis espacial de eventos delictivos se realizó considerando la información georreferenciada para Medellín y con base en técnicas estadísticas no paramétricas. En particular, se utilizó un KDE (Kernel Density Estimation), técnica que evita imponer supuestos sobre la distribución subyacente o proceso estocástico espacial que genera los datos asociados a cada uno de los delitos; además, capta muy ajustadamente la evolución de la densidad del delito durante el transcurso del tiempo.

La figura III.6-2 muestra la concentración espacial de los hurtos en el caso de Medellín, la cual se da mayormente en los centros de mayor actividad económica. Cabe destacar que, a pesar de haberse reducido de manera notable luego de las medidas de aislamiento social, el análisis espacial desarrollado indicó que la distribución espacial de los hurtos no cambió de manera notable durante la emergencia sanitaria ni mientras la vigencia del aislamiento social más estricto.

Figura III.6-2. Concentración espacial de hurtos a personas, semana 1 de 2020, Medellín



Fuente: tomado de informe final (DNP & CAOBA, 2020).

3.6.4.3. Análisis temporal del delito

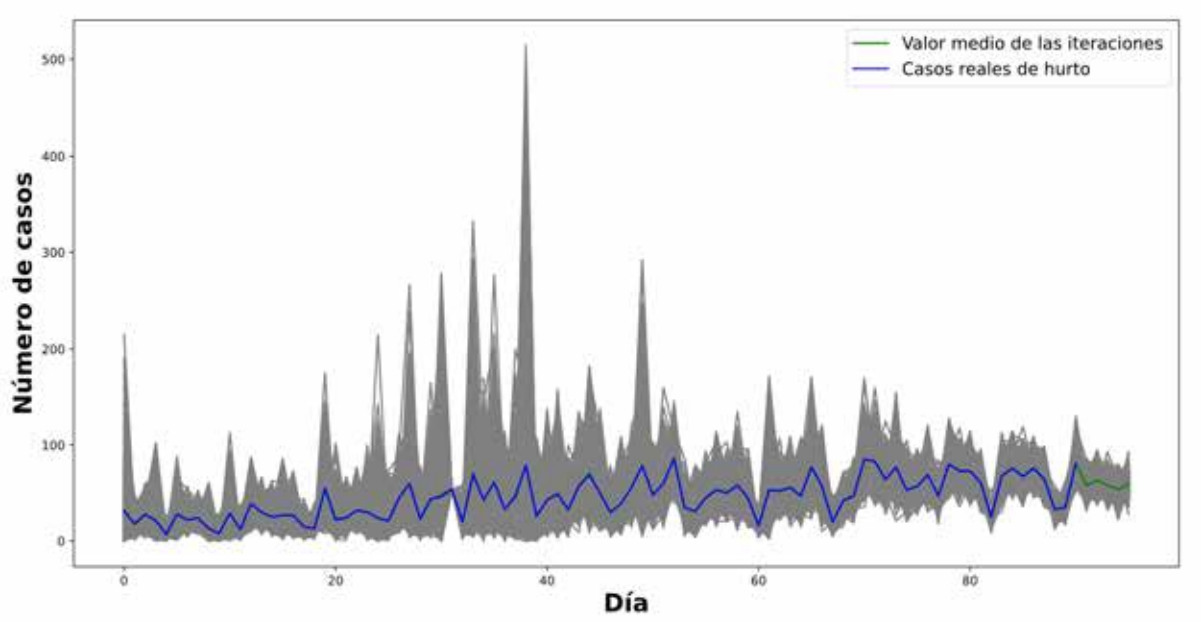
Se desarrolló un modelo de predicción de series de tiempo que relaciona la evolución de los indicadores de seguridad priorizados —homicidios, hurto, lesiones personales y violencia intrafamiliar— con la evolución de los casos de la COVID-19.

Algunos resultados de la predicción de hurtos pueden verse en la figura III.6-3, donde se representan con la línea azul los casos reales de hurto que se registrados en el periodo comprendido entre el 1 de abril y el 30 de junio de 2020, mientras que

la línea verde muestra la estimación *out of sample* de los casos de hurto, a través de la media muestral de 1000 realizaciones del proceso estocástico modelado. La estimación puntual (línea verde) es tan precisa que se solapa con la línea azul de los casos reales. Las correlaciones estimadas entre casos de contagio y eventos delictivos que se han identificado permiten calcular, con base en modelos de series de tiempo con ventanas móviles, estimaciones a corto plazo del número de eventos delictivos en términos de los casos de la COVID-19, como la que se observa en verde en el último tramo del período de análisis de la figura III.6-3.

Por supuesto, es importante tener en cuenta que la relación entre los indicadores de seguridad y el avance de la pandemia no necesariamente será estable en el tiempo, ya que en gran medida tal relación se debe a las consecuencias sobre la movilidad humana que tuvo la pandemia en sus diferentes etapas y también a las medidas de distanciamiento impuestas para mitigarla.

Figura III.6-3. Trayectorias del proceso estocástico para casos de hurto a personas ciudad de Medellín, 2020



Fuente: tomado de informe final (DNP & CAOBA, 2020).

Cabe destacar que los códigos desarrollados por los científicos de datos vienen con sus respectivos manuales de usuario, donde se especifica con detalle y de manera clara su usabilidad, incluso para quienes no son expertos en ciencia de datos y programación. Este tipo de materiales son los que proveen la posibilidad de réplica para otras ciudades en caso de disponer de la información necesaria.

3.6.4.4. Resultados

Tanto el análisis descriptivo como el espacial y el temporal arrojaron información útil para describir los patrones de los eventos delictivos observados durante 2020; a su vez, dicho análisis ayudó a diferenciar lo ocurrido durante ese año con patrones observados en un año más regular en términos de movilidad de personas, como fue 2019. Además, y dada la documentación adjunta al desarrollo del prototipo de análisis de datos georreferenciados de delito, el mismo tipo de análisis puede replicarse en otros contextos o ciudades donde se disponga de información de este tipo.

3.6.4.5. Recomendaciones

El gran valor informativo que puede extraerse desde datos que contengan etiquetas espaciales y temporales puede incrementarse aún más en caso de contar con mecanismos para compartir información delincuencial. Sin embargo, compartir información sobre hechos delictivos aun dentro de las propias entidades públicas no es una práctica generalizada y es una tarea compleja con varios desafíos que cubrir. No obstante, dado el valor que otorgan este tipo de análisis a las políticas de seguridad, resulta trascendental seguir invirtiendo en mecanismos eficientes y seguros para compartir información sobre eventos delictivos.

También se ha de resaltar que los patrones cambiantes de la movilidad de las personas, dados como consecuencia de la pandemia y las medidas de distanciamiento impuestas para contenerla, requieren que los modelos desarrollados en el presente proyecto se mantengan actualizados con la información más reciente que sea posible obtener.

CAPÍTULO 3.7

CONSIDERACIONES FINALES

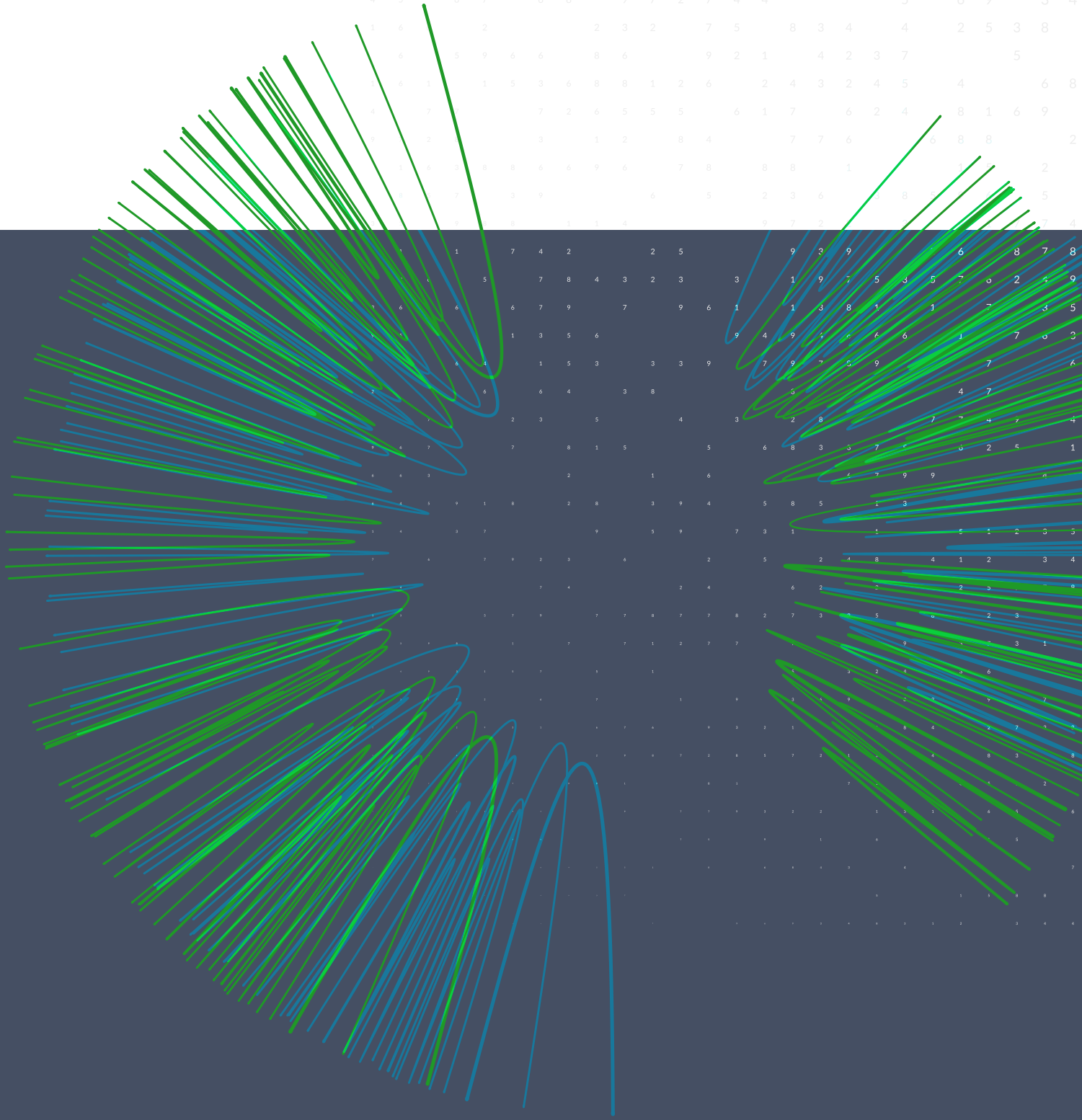
EL DESARROLLO DE LOS PROYECTOS DE ANALÍTICA, PRESENTADOS EN LOS SEIS APARTADOS ANTERIORES DE ESTA PARTE III, EVIDENCIAN LA UTILIDAD DE LA ANALÍTICA DE DATOS PARA ABORDAR PROBLEMÁTICAS DE INTERÉS PÚBLICO PARA LOS GOBIERNOS.

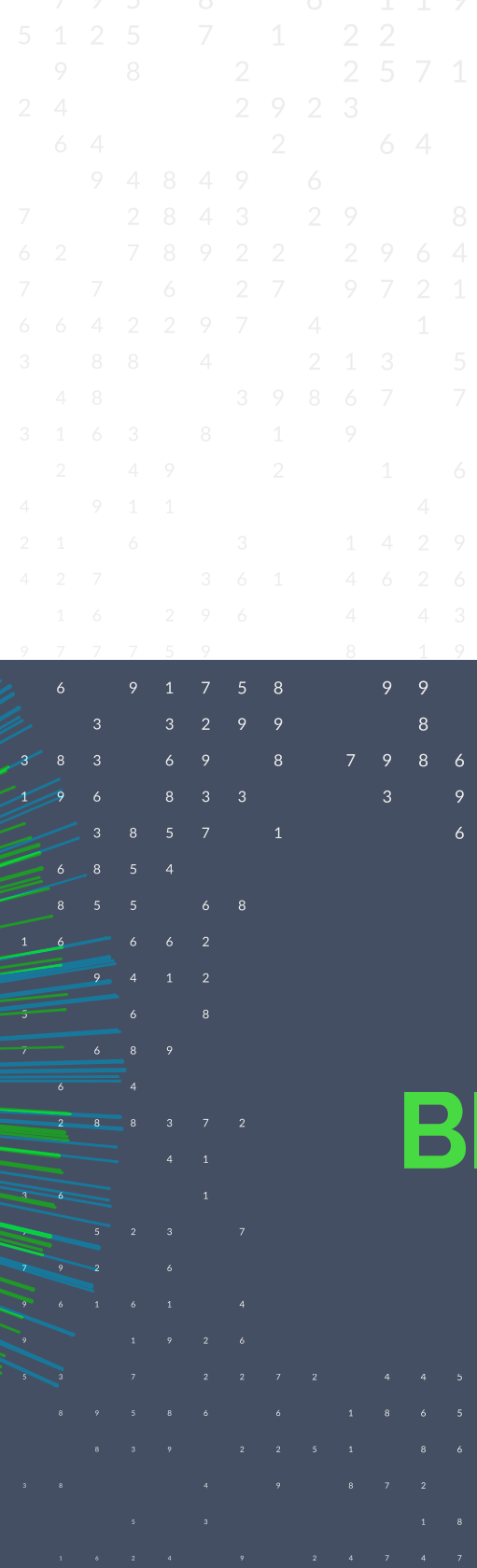
De manera general, los seis proyectos tuvieron una consecución exitosa de los objetivos planteados desde el inicio, gracias al trabajo continuo y conjunto de los científicos de datos de la Alianza CAOBA y los expertos de política pública de las direcciones técnicas del DNP. La elaboración de los proyectos dejó lecciones aprendidas en varios frentes: calidad y acceso a los datos, acompañamiento de los expertos técnicos en políticas públicas, seguimiento oportuno a las metas y objetivos de los proyectos, y utilización de resultados por parte de los tomadores de decisiones.

En cuanto a la disposición y acceso a datos de calidad, se identificó la importancia de emprender una gestión temprana con las entidades administradoras o custodias de los datos para elaborar los acuerdos de intercambio y los mecanismos técnicos necesarios para el intercambio de datos. Esa labor tiene el fin de disponer de los datos que se requieren al iniciar los proyectos, los cuales deben estar debidamente acompañados de un diccionario de datos, en el que se describan las variables y los campos para su interpretación adecuada.

El acompañamiento de los expertos en política pública también es fundamental para la elaboración pertinente de los proyectos. En el caso de Manos en la Data, los técnicos del DNP acompañaron de manera constante el trabajo elaborado por los científicos de datos. Se constituyó oportunamente una agenda de trabajo y se adelantaron reuniones periódicas de seguimiento. Esta práctica es muy positiva, dado que facilita subsanar fallas de información entre los equipos de trabajo, e identificar de manera conjunta estrategias o acciones que direccionen el proyecto, en caso en que haya alguna barrera técnica u organizativa.

Por último, se destaca la trascendencia de visualizar adecuadamente los resultados para la toma de decisiones posterior. En el caso de los seis proyectos de Manos en la Data, los resultados de los proyectos fueron socializados con diferentes públicos, entre los que se incluyen la ciudadanía, las direcciones técnicas directamente implicadas y las entidades externas al DNP pertenecientes al sector salud y de protección social y otros más. Este ejercicio es muy pertinente para continuar visibilizando en diferentes instancias la importancia de la analítica de datos para la toma de decisiones basadas en evidencia.





BIBLIOGRAFÍA

- Abdallah, Z., Du, L., & Webb, G. I. (2017). Data Preparation. En C. Sammut, & G. Webb, *Apoyo en la revisión de documentos relacionados con los sitios web del DNP.* Nueva York: Springer US.
- Amazon. (2020). *Amazon Machine Learning - evaluating models*. https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/ machinelearning-dg.pdf#evaluating_models
- Azevedo, A., & Santos, M. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conf. Data Mining.*
- Azzone, G. (2018). *Big data and public policies:Opportunities and challenges.* Milán: Department of Management, Economics and Industrial Engineerin.
- Banco Interamericano de Desarrollo. (2019). *Cuatro recomendaciones para promover políticas públicas basadas en evidencia.* https://blogs.iadb.org/efectividad-desarrollo/es/cuatro-recomendaciones-para-promover-politicas-publicas-basadas-en-evidencia/
- Banco Mundial. (2014). *World Bank. 2014. Central America: Big Data in Action for Development.* https://openknowledge.worldbank.org/handle/10986/21325
- Barbero, M.; Coutuer, J.; Jackers, R.; Moueddene, K.; Renders, E.; Stevens, W., Toninato, Y.; Van der Peijl, S. & Versteele, D. (2016). *Big data analytics for policy making.* [A study prepared for the European Commission DG INFORMATICS (DG DIGIT)]. https://joinup.ec.europa.eu/sites/default/files/document/2016-07/dg_digit_study_ big_data_analytics_for_policy_making.pdf
- Banco Interamericano de Desarrollo. (2020). *El gig data en los tiempos del coronavirus.* https://blogs.iadb.org/ideas-que-cuentan/es/el-big-data-en-los-tiempos-del-coronavirus/
- *Big Data* UN Global Working Group. (2019, 22 de febrero). *Big data solutions for enhancing tax compliance*. https://marketplace.officialstatistics.org/big-data-solutions-for-enhancing-tax-compliance
- *Big Data* UN Global Working Group. (2019, 13 de febrero). *Smart meter data potential for detecting unoccupied dwellings*. https://marketplace.officialstatistics.org/smart-meter-data-potential-for-detecting-unoccupied-dwellings
- *Big Data* UN lobal Working Group. (2019, 13 de febrero). *Integrated Environment System (IES): using environmental sensing systems and data analytics for real-time environmental information.* https://marketplace.officialstatistics.org/ integrated-environment-system-ies-using-environmental-sensing-systems-and-data-analytics-for-real-time-environmental-information

- Brownlee, J. (2020). *Tour of Evaluation Metrics for Imbalanced Classification. Machine Learning Mastery*. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

- CAF-banco de desarrollo de América Latina-. (2019). *Ciencia de datos para mejorar las políticas públicas: La experiencia de Córdoba*. https://www.caf.com/es/actualidad/noticias/2019/03/ciencia-datos-para-mejorar-las-politicas-publicas-la-experiencia-en-cordoba/

- Center for Leading Innovation & Collaboration. (2020). *The National COVID Cohort Collaborative (N3C): A National Data Sharing Partnership to Fight COVID-19*. https://clic-ctsa.org/posters/national-covid-cohort-collaborative-n3c-national-data-sharing-partnership-fight-covid-19

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). The CRISP-DM user guide. *4th CRISP-DM SIG* [Workshop in Brussels in March, 1999].

- Comisión Europea. (2016). *Big data Analytics for policy making*. https://joinup.ec.europa.eu/ sites/default/files/document/2016-07/dg_digit_study_big_data_analytics_for_ policy_making.pdf

- Departamento Administrativo de la Función Pública. (2016). *Modelo Integrado de Planeación y Gestión*. https://secretariageneral.gov.co/sites/default/files/generalidades_mipg.pdf

- Datereportal. (2020). *Digital 2020: Panorama Digital Global*. https://datareportal.com/ reports/digital-2020-global-digital-overview

- Departmento Nacional de Planeación. (2018, 17 de abril). *Política Nacional de Explotación de Datos [*Big Data*]*. (Documento *CONPES 3920* DNP. https://colaboracion.dnp.gov.co/ CDT/Conpes/Econ%C3%B3micos/3920.pdf

- Departamento Nacional de Planeación. (2019, 8 de noviembre). *Política Nacional por la Transformación Digital e Inteligencia Artificial*. (Documento CONPES 3975). DNP. https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3975.pdf

- Departamento Nacional de Planeación. (2021, 11 de febrero). *Política para la reactivación, la repotenciación y el crecimiento sostenible e incluyente: nuevo compromiso por el futuro de Colombia* (Documento CONPES 4023). DNP. https://colaboracion.dnp.gov.co/ CDT/Conpes/Econ%C3%B3micos/4023.pdf

- Departamento Nacional de Planeación & Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics*. (2020). *Informe CAF - DNP. Sistema Nacional de Cuidado*. DNP - Alianza CAOBA.

- Departamento Nacional de Planeación & Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics*. (2020). *Informe final. Caracterización Zonas COVID-19*. DNP - Alianza CAOBA.

- Departamento Nacional de Planeación & Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics*. (2020). *Informe final. Matriz Origen destino Manos en la data*. DNP - Alianza CAOBA.

- Departamento Nacional de Planeación & Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics*. (2020). *Informe final. Dinámica del empleo formal y las empresas bajo la contingencia del COVID-19*. DNP - Alianza CAOBA.

- El Tiempo. (2019). *Fórmulas médicas polémicas que todos pagamos*. https://www.eltiempo.com/salud/formulas-medicas-polemicas-que-todos-pagamos-396330

- Foro Económico Mundial. (2019). *¿Cuantos datos se generan cada día?* https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

- Foro Económico Mundial. (2020). *Tres formas en las que COVID-19 está transformando la analítica avanzada y la inteligencia artificial.* https://www.weforum.org/agenda/ 2020/07/3-ways-covid-19-is-transforming-advanced-analytics-and-ai/

- Grupo Asesor de Expertos Independientes sobre la Revolución de los Datos para el Desarrollo Sostenible –GAEI–. (2014). *Un mundo que cuenta*. https://repositorio.cepal.org/ bitstream/handle/11362/37889/UnMundoqueCuenta.pdf?sequence=1&isAllowed=y

- Govtech Singapore. (2016). *Rogue Train: A big data Story*. Technews: https://www.tech.gov.sg/media/technews/rogue-train-a-big-data-story

- Headey *et al*., (2020). *Impacts of COVID-19 on childhood malnutrition and nutrition-related mortality:* https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31647-0/fulltext

- International Data Corporation –IDC–. (2020, mayo). *El Pronóstico Global de DataSphere de IDC muestra un crecimiento continuo y constante en la creación y consumo de datos*. https://www.idc.com/getdoc.jsp? containerId=prUS46286020

- Infométrika. (2020). *Modelo de explotación de datos para las entidades públicas*. Bogotá.

- Instituto Colombiano de Bienestar Familiar. (2015). *Documento general de análisis Encuesta Nacional de la Situación Nutricional en Colombia - ENSIN 2015*. ICBF. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/libro-ensin-2015.pdf

- Lean-Data. (2018). *Here is how to start with data quality*. https://www.lean-data.nl/data-quality/here-is-how-to-start-with-data-quality/

- Li, M., Hong, F., Sun, R., & Che, C. (2016). *International Conference on Communications, Information Management and Network Security.* The Application of *Big Data* Analysis Techniques and Tools in Intelligence Research.

- Marchi, G., Lucertini, G., & Tsoukias, A. (2016). From Evidence-Based Policy-Making to Policy Analytics. *Annals of Operations Research, Springer Verlag*, pp.15-38. From Evidence-Based Policy-Making to Policy Analytics.

- Mattick, K., Johnston, J., & Croix, A. D. (2018). How to…write a good research question. *The Clinical Teacher*, 104-108. doi:doi:10.1111/tct.12776

- McKinsey Global Institute. (2016). *The age of analytics: Competing in a data-driven world*. https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.pdf

- Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

- Mohamed, M., & Weber, P. (2020). *School of Engineering and Applied Science*. Trends of digitalization and adoption of big data & analytics among UK SMEs: https://arxiv.org/ftp/arxiv/papers/2002/2002.11623.pdf

- Moore, M. (1995). *Creating Public Value: Strategic Management in Government*. Cambridge, Harvard.

- Ontiveros, E.; Vizcaíno, D. & López, V. (2017, marzo). *Las ciudades del futuro: inteligentes, digitales y sostenibles.* Fundación Telefónica - Ariel.

- Organización para la Cooperación y el Desarrollo Económicos. (2014). *Revisión de Gobierno Digital de Colombia: hacia un sector público impulsado por el ciudadano*. https://www.oecd.org/gov/digital-government/Digital%20Gov%20Review%20Colombia%20[Esp]%20def.pdf

- Organización para la Cooperación y el Desarrollo Económicos. (2015). *Big data for growth and well being*. https://read.oecd-ilibrary.org/science-and-technology/data-driven-innovation_9789264229358-en#page45

- Organización para la Cooperación y el Desarrollo Económicos. (2018). *Revisión del gobierno digital en Colombia. Utilización estratégica de los datos en el sector público de los datos*. https://read.oecd-ilibrary.org/governance/revision-del-gobierno-digital-en-colombia/utilizacion-estrategica-de-datos-en-el-sector-publico-colombiano_9789264292147-6-es#page10

- Organización para la Cooperación y el Desarrollo Económicos. (2020). *Smart Cities and Inclusive Growth*. http://www.oecd.org/cfe/cities/OECD_ Policy_Paper_Smart_Cities_and_Inclusive_Growth.pdf

- Organización de las Naciones Unidas —ONU Mujeres—. (s.f.). *ODS 8: Promover el crecimiento económico sostenido, inclusivo y sostenible, el empleo pleno y productivo y el trabajo decente para todas y todos*. https://www.unwomen.org/es/news/in-focus/women-and-the-sdgs/sdg-8-decent-work-economic-growth

- Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research, 12*, 2825-2830.

- Press, G. (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#aef2c226f637

- *Produraduría General de la Nación*. (2015). *El valor público, una nueva visión de la gestión*. https://www.procuraduria.gov.co/sipre/media/file/Publicaciones/6/Valor%20Publico.pdf

- Rodríguez, P., Palomino, N., & Mondaca, J. (2017). *Using big data and its Analytical Techniques for Public Policy Design and Implementation in Latin America and the Caribbean*.

- Rodríguez, P., Palonimo, N., & Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas para el diseño e mplementación de políticas públicas en Latinoamérica y el Caribe*. https://publications.iadb.org/publications/spanish/document/El-uso-de-datos-masivos-y-sus-t%C3%A9cnicas-anal%C3%ADticas-para-el-dise%C3%B1o-e-implementaci%C3%B3n-de-pol%C3%ADticas-p%C3%BAblicas-en-Latinoam%C3%A9rica-y-el-Caribe.pdf

- Rodríguez, P., Palonimo, N., & Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas para el diseño e mplementación de políticas públicas en Latinoamérica y el Caribe*. https://publications.iadb.org/publications/spanish/document/El-uso-de-datos-masivos-y-sus-t%C3%A9cnicas-anal%C3%ADticas-para-el-dise%C3%B1o-e-implementaci%C3%B3n-de-pol%C3%ADticas-p%C3%BAblicas-en-Latinoam%C3%A9rica-y-el-Caribe.pdf

- Ryan , M., & Talabis, M. (2014). *Information Security Analytics*. https://www.sciencedirect.com/topics/computer-science/text-mining

- Schöllhammer, R., Parycek, P., & Höchtl, J. (2016). *Big data in the policy cycle: Policy decision making in the digital era*. https://www.tandfonline.com/doi/full/10.1080/ 10919392.2015.1125187

- Scikit-Learn Developers. (2016). *Choosing the right estimator.* https://scikit-learn.org/stable/ tutorial/machine_learning_map/index.html

- Soric, D., Stajcer, M., & Orescanin, D. (2017). Análisis eficiente de redes sociales en arquitecturas de big data, [40a Convención Internacional de 2017 sobre Tecnología de la Información y la Comunicación, Electrónica y Microelectrónica (MIPRO)]. https://ieeexplore.ieee.org/document/7973640

- Statista. (Mayo de 2020). *Volumen de datos / información creado en todo el mundo desde 2010 hasta 2024*. https://www.statista.com/statistics/871513/worldwide-data-created/

- Studinka, J., & Guenduez, A. (2019). *The Use of Big Data in the Public Policy Process: Paving the Way for EvidenceBased Governance*. https://www.alexandria.unisg.ch/255680/1/ Studinka%20and%20Guenduez%20-%20The%20Use%20of%20Big%20Data%20in%20the%20Public%20Policy%20Process-%20Paving%20the%20Way%20for%20Evidence-Based%20Governance%20.pdf

- Swalin, A. (2018). *Choosing the Right Metric for Evaluating Machine Learning Models - Part 1*. Medium.com: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

- Swalin, A. (2018). *Choosing the Right Metric for Evaluating Machine Learning Models - Part 2*. KDnuggets: https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

- Targio Hassem, I., Yaqoob, I., & Badrul Anuar, N. (2014). *The rise of big data on cloud computing: Review and open research issues*: https://www.researchgate.net/ publication/264624667_The_rise_of_Big_Data_on_cloud_computing_Review_and_open_research_issues

- Workera. (s.f.). *AI Career Pathways: Put yourself on the Right Track.*


El futuro
es de todos
DNP
Departamento
Nacional de Planeación


BANCO DE DESARROLLO
DE AMÉRICA LATINA


El futuro
es de todos
DNP
Departamento
Nacional de Planeación


BANCO DE DESARROLLO
DE AMÉRICA LATINA