CONSEJO ASESOR TÉCNICO DEL SISTEMA ESTADÍSTICO NACIONAL -CASEN

SALA ESPECIALIZADA DE MODERNIZACIÓN TECNOLÓGICA DE LA PRODUCCIÓN ESTADÍSTICA

DOCUMENTO DE RECOMENDACIONES

LÍNEAS DE INVESTIGACIÓN:
BIG DATA E INTEROPERABILIDAD
SEGURIDAD DE LA INFORMACIÓN

Diciembre de 2022

Carrera59 No. 26 – 70 Interior 1, CAN
Bogotá, Colombia
Teléfono (+57-1) 597 8300 ext. 3210-3294

DEPARTAMENTO ADMINISTRATIVONACIONAL DE ESTADÍSTICA -DANE

BEATRIZ PIEDAD URDINOLA CONTRERAS
Director

LEONARDO TRUJILLO OYOLA
Subdirector

MARÍA FERNANDA DE LA OSSA ARCHILA
Secretaria General

DIRECTORES TÉCNICOS

ANDREA RAMÍREZ PISCO
Dirección de Metodología y Producción Estadística

ANGELA PATRICIA VEGA LANDAETA
Dirección de Censos y Demografía

JULIETH ALEJANDRA SOLANO VILLA
Dirección de Regulación, Planeación, Estandarización y Normalización

JUAN PABLO CARDOSO TORRES
Dirección de Síntesis y Cuentas Nacionales

SANDRA LILIANA MORENO MAYORGA
Dirección de Geoestadística

FABIÁN RICARDO MEJÍA OSPINA
Dirección de Recolección y Acopio

MAURICIO ORTIZ GONZÁLEZ
Dirección de Difusión y Cultura Estadística

VALÉRIE GAUTHIER UMAÑA
LEÓN DARÍO PARRA
Expertos
Consejo Técnico Asesor del Sistema Estadístico Nacional -CASEN

MÓNICA PATRICIA PINZÓN TORRES
Coordinadora de Planificación y Articulación Estadística

SOFÍA SÁNCHEZ GRANADOS
RUTH CONSTANZA TRIANA ACUÑA
Coordinación de Planificación y Articulación Estadística

CONTENIDO

Introducción

El Departamento Administrativo Nacional de Estadística -DANE, a partir de la Ley 1955 de 2019 en su artículo 155 se constituye en el ente rector, coordinador y regulador del Sistema Estadístico Nacional - SEN y con el Decreto reglamentario 2404 de 2019, establece la conformación del Consejo Asesor Técnico del Sistema Estadístico Nacional -CASEN, la sala general y sus salas especializadas en diferentestemáticas: i) Sala especializada para la modernización tecnológica de la producción estadística, ii) Sala especializada de salud, bienestar social y demografía, iii) Sala especializada de gobierno, seguridad y justicia, iv) Sala especializada de geografía, medio ambiente y ordenamiento territorial y v) Sala especializada de economía. Cada sala especializada estará integrada por tres (3) expertos.

Este documento tiene como objetivo presentar al SEN y a la sala general del CASEN, las recomendaciones realizadas para el año 2022 por los miembros de la Sala Especializada de Modernización Tecnológica de la Producción Estadística y el DANE.

Durante el año 2022, se abordaron las siguientes dos líneas de investigación: i) Big Data e Interoperabilidad y ii) Seguridad de la Información.

Se resalta la metodología alterna que se utilizó para el desarrollo de la sala:

Reuniones plenarias: en donde los pares del DANE (Oficina de Sistemas), presentaron diferentes temas de interés en el contexto de la entidad y las consecuentes preguntas, con el objetivo de recibir realimentación por parte de los expertos. Se destacan las reuniones realizadas sobre BigData y Blockchain.

1. Reuniones internas: las cuales contaron con el acompañamiento de invitados, que nutrieron el debate y aportaron desde sus campos del conocimiento a ampliar la mirada del DANE, en temas como la seguridad de los datos en la nube desde el aspecto jurídico y de ciberseguridad.

2. Trabajo bilateral con los expertos: reuniones realizadas con el objetivo de hacer seguimiento al cronograma de actividades de la sala y a los compromisos adquiridos por cada experto. Se destaca la articulación para lograr el acompañamiento a diferentes proyectos del DANE: el documento final del paper índice de noticias- 2022 "Colombian agricultural sector's early estimator of GDP post-pandemic COVID19 using Google News and Google Trends", las

recomendaciones metodológicas y sobre los resultados al proyecto de cálculo de indicador 16 del Objetivo de Desarrollo Sostenible -ODS, 16 usando redes sociales; y la participación en la conferencia de Matemáticas aplicadas e industriales – MAPI, que se llevó a cabo en la ciudad de Medellín del 8 al 10 de junio de 2022.

A continuación, se presenta a la sala general del CASEN los resultados del trabajo realizado en las dos líneas de investigación. Se inicia con el planteamiento de los objetivos del documento, se continúa con los antecedentes, el marco conceptual de la sala y finaliza con las recomendaciones de los expertos para las líneas de interés.

1. Objetivos

1.1 Objetivo general

Presentar a la sala general del CASEN, el trabajo que realizó la Sala Especializada de Modernización Tecnológica de la Producción Estadística durante 2022, en las dos líneas de investigación: i) Big Data e interoperabilidad y ii) Seguridad en la Información. Además, socializar las recomendaciones esbozadas por los expertos durante las reuniones para el fortalecimiento y uso de la información estadística.

1.2 Objetivos específicos

- Promover la articulación entre las entidades del SEN para la transferencia de datos y el desarrollo de ejercicios piloto de intercambio de información.
- Fortalecer los espacios de diálogo entre las entidades públicas y la academia, generando modelos de aprendizaje colaborativo.
- Revisar y considerar marcos metodológicos internacionales para el mapeo de necesidades y capacidades tecnológicas regionales y locales que permitan la interoperabilidad.
- Realimentar metodológicamente los proyectos adelantados por el DANE en materia de producción de estadísticas experimentales.
- Reconocer el marco normativo colombiano asociado a la seguridad de la información y a la protección de datos en la nube.

2. Antecedentes

Durante los dos primeros años de gestión de la Sala de Modernización Tecnológica de la Producción Estadística del CASEN (2020- 2021), se analizaron diferentes iniciativas relativas a la implementación de nuevas tecnologías al SEN. Se destacan como principales logros, los siguientes:

Principales logros **Sala de Modernización Tecnológica de Modernización Tecnológica- CASEN (2020- 2021)**	
Año 2020	**Año 2021**
➢ Realización del seminario Web "Desafíos del SEN para la interoperabilidad de datos e información estadística". ➢ Comentarios y recomendaciones al proceso de interoperabilidad liderado por el Ministerio de Tecnologías de la Información y las Comunicaciones. ➢ Identificación de los niveles del DANE en los dominios del Modelo de madurez del Marco de Interoperabilidad: dominio organizacional, político-legal, semántico y técnico. ➢ Aportes al registro "Planilla Integrada de Liquidación de Aportes – PILA" como caso exitoso de interoperabilidad. ➢ Documento de recomendaciones 2020.	➢ Asesoría al "Proyecto Comité de Administración de Datos -CAD" en los temas de: seguridad, gestión de proveedores, lenguaje común, roles y en la definición de los principios éticos. ➢ Recomendaciones a los proyectos de: desarrollo y distribución de app web para la recolección de datos, Web APP y el índice de noticias, educación D4N, anonimización –CEED y DSCN sector carbón y cruces difusos en PILA y DEST, la detección de anomalías - Censo Económico y la automatización de procesos de calidad- Temática e Imputación de datos con cruce difuso- Gran Encuesta Integrada de Hogares -GEIH. ➢ Presentaciones sobre: política de gobierno de registros y fuentes alternativas, el modelo de gestión de datos, la política de seguridad de la información DANE, la anonimización del censo de edificaciones, y, sobre criptografía: sus implicaciones, perspectivas y conceptos básicos. ➢ Documento de recomendaciones 2021.

A continuación, se presentan los antecedentes de cada una de las líneas de investigación abordadas durante el año 2022.

2.1 Línea Big Data e Interoperabilidad

Big Data

Las Naciones Unidas definieron el Big Data como un volumen masivo de datos, estructurados y no estructurados, que son demasiado grandes y difíciles de procesar con bases de datos y software tradicionales (2012). Según eso, y a pesar que Colombia fue uno de los primeros países en la región en adoptar una política de Big Data, es claro que aún hay un largo camino por recorrer para explorar efectivamente esta posibilidad favoreciendo su desarrollo. Sobre todo, porque el objetivo es tener personas y organizaciones (públicas y privadas) con la capacidad de generar bienes y servicios a través del uso de datos y que los tomadores de decisiones de las organizaciones reconozcan el valor de sus acciones basadas en datos para aumentar la eficiencia y la productividad.

Teniendo en cuenta lo anterior, la línea de Big Data que se trabaja en la Sala Especializada de Modernización Tecnológica tiene como objetivo recomendar y ampliar prácticas de uso de manejo de datos en el DANE.

Por ejemplo, desde la Sala de Modernización Tecnológica se brindó realimentación al proyecto "Desarrollo y distribución de aplicaciones web para la recolección de datos" que ha adelantado el DANE, el cual tiene por objetivo desarrollar y distribuir, por medio de una red social, una aplicación web, como método alternativo de recolección de información estadística oficial. Su alcance se enfoca en desarrollar una aplicación en versión beta o PMV para validar las ventajas y viabilidad de escalar el método de recolección a otros productos de información del DANE. Prueba en un caso de recolección de datos para un indicador de Desarrollo Sostenible, tomando el indicador 16.b.1 y el 10.3.1 de la proporción de la población que declara haberse sentido personalmente víctima de discriminación o acoso en los 12 meses anteriores por motivos de discriminación prohibidos por el derecho internacional de los derechos humanos; así como validar las ventajas de distribuir la aplicación en la red social Facebook(Cobertura, Costos).

En esta medida, el proyecto "Desarrollo y distribución de aplicaciones web para la recolección de datos" se plantea teniendo en cuenta que la recolección de información por encuestas es cada vez un reto mayor para los INEs. Por esta razón, se ha identificado como una oportunidad para la recolección de datos el hecho que las personas gasten cada vez más tiempo en el uso de dispositivos electrónicos, redes sociales y gran parte del contenido de internet lo generan los usuarios de la Unidad Generadora de Datos -UGD. Por otro lado, los algoritmos de las plataformas de redes sociales permiten la distribución rápida ydirigida

de contenido, en donde la obtención de valor es la motivación del usuario para suministrar información (usar una app) y la experiencia de uso de una aplicación es más atractivo para el usuario que diligenciar un formulario.

Interoperabilidad

De conformidad con la definición adoptada por el Ministerio de Tecnologías de la Información en el Marco de interoperabilidad para Gobierno Digital, la Interoperabilidad es la capacidad de las organizaciones para intercambiar información y conocimiento en el marco de sus procesos de negocio para interactuar hacia objetivos mutuamente beneficiosos, con el propósito de facilitar la entrega de servicios digitales a ciudadanos, empresas y a otras entidades, mediante el intercambio de datos entre sus sistemas TIC (2019).

En el decreto 2404 de 2019 establece en su artículo 2.2.3.5.1 lo referente al intercambio de microdatos, registros administrativos y fuentes alternativas para la producción de estadísticas oficiales. Este artículo dictamina el intercambio de información estadística refiriéndose a las instituciones gubernamentales "En concordancia con lo anterior, las partes formalizarán el intercambio mediante acuerdos bipartitos en los cuales se hagan explícitas las condiciones de traslado de reserva que la legislación vigente permita. En los acuerdos de intercambio se privilegiarán las condiciones de protección de datos y de seguridad de la información del custodio del microdato a fin salvaguardar los riesgos sobre activos de información identificados por este...". Teniendo en cuenta el artículo mencionado, desde el año 2021, la Sala de Modernización Tecnológica de la Producción Estadística busca asesorar al SEN para proponer un esquema para el intercambio de datos entre los miembros del sistema, que garanticen la seguridad de la información, así como el uso ético y adecuado.

Durante los años 2021 y 2022, los expertos de la sala trabajaron sobre la propuesta del DANE para la implementación del Sistema Nacional de Interoperabilidad para la gestión de la información estadística entre las diferentes entidades productoras.

El primer tema que se recomendó a la sala, fue que el Estado avanzara en su modelo de interoperabilidad a través de su correspondiente plataforma X-ROAD, en la medida que articule y apalanque su desarrollo digital en tres pilares fundamentales: 1. agilidad en los procesos de estandarización, normalización e intercambio de datos, 2. eficiencia en el uso de los recursos e integración de la información, y 3 intercambio de información protegiendo en todo momento los datos personales de los usuarios y ciudadanos adscritos al sistema. Se debe tener en cuenta a su vez, desde el lado de la oferta, aspectos como: el grado de conectividad de las instituciones y su nivel de madurez digital, la colaboración y grado de apertura para el trabajo interinstitucional, los procesos de gestión del cambio implementados en cada

organismo, y el respectivo marco de gobernanza que dará las pautas para dirigir el sistema (BID, 2019).

Además, la sala resaltó la importancia de implementar los marcos definidos por el Ministerio de Tecnologías de la Información y las Comunicaciones entre los miembros del SEN, el intercambio de información; así como recomendar modelos de seguridad y control de la información del DANE y explorar y recomendar modelos de desidentificación/ anonimización de la información.

2.2 Línea Seguridad de la Información

La seguridad de la información puede definirse como el conjunto de medidas preventivas y de reacción que permiten resguardar y proteger la información. Es decir, son todas aquellas políticas de uso y métodos que afectan al tratamiento de los datos que se utilizan en una entidad. Teniendo en cuenta que la seguridad de la información es una pieza fundamental para que las organizaciones que manejan datos puedan llevar a cabo sus operaciones sin asumir demasiados riesgos, puesto que los datos que se manejan son esenciales para la razón social de la entidad, sobre todo en el caso de los INEs, que se encargan de procesar y analizar la información de todo un país de manera trasversal. Se exponen los ejemplos internacionales de países como Estados Unidos y Australia donde en sus respectivas oficinas de estadística presentan lineamientos como "...para salvaguardar la información incluyen un cifrado sólido de los datos, acceso restringido según la necesidad y supervisión de todo el personal, incluidas auditorías periódicas. Después de la recopilación y el procesamiento de datos, el (Australian Bureau of Statistics) ABS elimina los nombres y direcciones de otra información personal y del hogar" (ABS, 2021). Para el caso de Estados Unidos se encuentra que tienen respecto a la seguridad de la tecnología que usan cuatro partes fundamentales:

- Diseño de sistemas seguros: trabajando con socios de la industria, diseñan sistemas con muchas capas de seguridad para defenderse y neutralizar las amenazas cibernéticas.
- Recopilación segura de datos: han desarrollado y mantienen una conexión segura a Internet para recopilar su información.
- Datos cifrados: sus datos se cifran durante la recopilación de datos y luego se almacenan en la red privada e interna de la Oficina del Censo, que está aislada de Internet por cortafuegos y otras medidas de seguridad.
- Acceso limitado: los sistemas de datos están protegidos por autenticación de dos factores.

Asimismo, utilizan medidas de seguridad de vanguardia para proteger la identidad de los habitantes basados en tres pilares fundamentales:

- Solo con fines estadísticos: no identifican a las personas en los datos que publicamos. Solo publican estadísticas.
- Garantía de confidencialidad: las políticas y salvaguardas estadísticas ayudan a garantizar la confidencialidad de la información.
- Estándares de confidencialidad: Cuentan con una Junta de Revisión de Divulgación que verifica que cualquier producto de datos que divulguen cumpla con los estándares de confidencialidad.

Uno de los grandes retos de la sociedad de la información es la salvaguarda de los datos como su principal recurso. En este contexto, las instituciones tanto públicas como privadas deben hacer frente al reto de gestionar y garantizar la seguridad informática de los diferentes sistemas de información y bases de datos que manejan. No solo porque la información y los datos se han convertido en el principal activo de las organizaciones, sino también por los derechos de privacidad y habeas data de relativos a la información personal a los cuales son sujetos todos los ciudadanos.

Teniendo el anterior referente, la sala de modernización tecnológica durante el año 2022, discutió los retos y desafíos a los cuales se enfrenta el sistema estadístico nacional en materia de seguridad informática haciendo énfasis tanto en la importancia que guarda el tener la tecnología adecuada para responder de manera eficaz ante las diferentes amenazas que podrían vulnerar la salvaguarda de la información personal de los usuarios, como el marco normativo y las leyes que protegen a los individuos en relación a sus datos personales. En ese sentido, se realizaron varios conversatorios con expertos técnicos y jurídicos quienes dieron luz sobre las variables y derroteros para tener en cuenta en la implementación de sistemas de seguridad informática integrales.

3. Marco conceptual

Los conceptos presentados a continuación son relevantes para las líneas de investigación y se relacionan con los temas tratados durante la vigencia dentro de la sala:

Aprendizaje de máquinas: es un método de análisis de datos que automatiza la construcción de modelos analíticos. Es una rama de la inteligencia artificial basada en la idea que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.

Blockchain: es un sistema de almacenamiento de información libre de falsificaciones. Es decir, una cadena de bloques que permite la trazabilidad sin intermediarios. Cuenta con tres características principales: i) Tal como un libro contable, cuenta con páginas enumeradas. Por lo que cobra sentido en conjunto y no puede faltar ninguna, ii) No se puede cambiar el contenido y iii) La información puede estar cifrada o no.

Gobierno de datos: es un sistema de toma de decisiones y responsabilidades para procesos relacionados con datos. La toma de decisiones es ejecutada de acuerdo con los modelos que describen quién puede tomar acciones, con qué datos y cuándo, así como las situaciones y los métodos para llevarlo a cabo.

Inteligencia artificial: se refiere a sistemas o máquinas que imitan la inteligencia humana para realizar tareas y pueden mejorar iterativamente a partir de la información que recopilan.

Información sensible: datos personales que revelan origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o a la vida sexual.

Resolución 0451 de 2020: en esta resolución se establece el funcionamiento del Comité de Administración de Datos -CAD, se dictamina que las funciones principales del CAD son: 1) Fomentar las colaboraciones en torno a los datos entre las entidades que conforman el Sistema Estadístico Nacional -SEN, apoyadas en la confianza y trabajo articulado de los equipos técnicos de las entidades involucradas, para el desarrollo de políticas públicas. 2) Establecer el esquema de gobernanza para evaluar los requerimientos de articulación de información estadística con el ciclo de las políticas públicas que se presenten para revisión del Comité. Lo anterior, con base en lo establecido en el artículo cuarto de la presente Resolución y demás aspectos que consideré pertinentes incluir. 3) Definir el marco ético que orientará la revisión de los casos, promoviendo el mayor uso y re -uso de los datos, garantizando que éste sea adecuado, responsable y seguro.

4. Recomendaciones

Luego de la Pandemia del COVID-19, la sociedad en su conjunto apropió la necesidad de adoptar las

nuevas tecnologías como parte de su diario vivir. En este contexto, el plan de modernización tecnológica liderado desde el anterior gobierno, que comenzó por la plataforma de interoperabilidad X-ROAD, hasta las diferentes iniciativas que ha implementado el DANE, para probar el uso de tecnologías asociadas al Big Data y la seguridad informática, son prueba del interés por parte del estado colombiano en seguir adelante con la ruta de modernización tecnológica del SEN.

No obstante, aún quedan grandes retos por resolver para lograr que el Sistema Estadístico Nacional sea competitivo y eficiente, en comparación con otros sistemas a nivel global. A continuación, las principales recomendaciones generales que deberá asumir la presente administración en la materia:

- Dar continuidad a las diferentes iniciativas que se desarrollaron en el CASEN durante la vigencia 2020 – 2022, y particularmente las relacionadas con el proceso de modernización tecnológica del SEN.

- La generación de espacios de interacción entre los diferentes actores y usuarios del SEN es aún un desafío por resolver. Involucrar en la charlas y sesiones que se realicen en cada sala a otros expertos o invitados de otras entidades públicas u organismos del Estado que sirvan como validadores tempranos desde la perspectiva del usuario de los proyectos que se encuentran en curso.

- Continuar trabajando de la mano con el sector académico y otras entidades públicas para profundizar la cooperación en investigación, formación y proyectos de extensión que puedan beneficiar tanto a la sociedad civil como al DANE. En esta medida, es importante concretar los convenios bilaterales como el que se encuentra en trámite con la Universidad EAN u otras universidades, con el fin de ampliar la cooperación e integrar equipos multidisciplinarios que le aporten a las iniciativas lideradas por el DANE.

- En materia de articulación, resalta la importancia de generar mayores espacios de interacción y trabajo colectivo entre las diferentes salas que integran el CASEN, con el fin de estructurar e implementar iniciativas que sean transversales a los temas de salud, economía, política, medio ambiente y modernización tecnológica.

- Aunado a lo anterior, se hace necesario crear nuevos espacios para que las salas del CASEN interactúen con los diferentes organismos e instituciones que producen información a nivel nacional, departamental y municipal, y que de allí se generen mesas de concertación para la creación de nuevos proyectos que fortalezcan el SEN y su articulación con el resto de la economía.

4.1 Big Data e Interoperabilidad

Big Data

En el marco de las reuniones desarrolladas en la presente sala durante el año 2022, el Departamento Nacional de Estadística -DANE presentó ante los expertos una serie de proyectos orientados a la modernización tecnológica para la producción y gestión de información estadística con el objetivo de recibir realimentación para continuar avanzando en la implementación de nuevas fuentes de información. Los proyectos se centraron en la producción de indicadores a través del uso de fuentes alternativas de información para la toma de decisiones basadas en evidencia.

A continuación, se presentan las principales recomendaciones realizadas por los expertos a cada uno de los proyectos: 1) Cálculo de indicadores del ODS 16 usando redes sociales, 2) Índice de noticias como estimador temprano de la actividad económica y 3) Desarrollo y distribución de Aplicaciones Web para la recolección de datos.

Proyecto de "Cálculo de indicadores del ODS 16 usando redes sociales"

Durante el primer semestre de 2022, se presentó el proyecto de cálculo de indicadores pertenecientes al Objetivo de Desarrollo Sostenible -ODS 16 usando redes sociales, que tiene por objetivo: "Obtener mediciones complementarias de los indicadores 16.b.1 y 16.7.2, así: 1. Porcentaje de usuarios de Facebook, por sexo, con comentarios que incluyen lenguaje discriminatorio (trato diferente, rechazo o maltrato). 2. Usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la toma de decisiones inclusiva. 3. Usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la toma de decisiones receptiva" (DANE, 2022).

Partiendo de lo anterior, se planteó una metodología exploratoria en la cual se utilizan métodos de Web Scraping y modelos de aprendizaje de máquinas, supervisado y no supervisado, que permitan mapear y discriminar la percepción de la población en las redes sociales acerca del tema de discriminación o que apunte al objetivo 16 de desarrollo sostenible. En este sentido, es válido generar lineamientos para la clasificación y selección de la información, así como la escogencia de las fuentes a partir de las cuales se extraen los datos.

Recomendaciones de los expertos y sugerencias al proyecto de "Cálculo de indicadores del ODS 16 usando redes sociales"

Las siguientes recomendaciones hacen referencia al aspecto metodológico del proyecto:

1. Ampliar la técnica de Web Scraping a otras redes sociales, segmentando por grupos etarios, según redes de predilección entre cada grupo, para ello se podría complementar la técnica con:

   - Web Scraping a Blogs u otras fuentes que hablen sobre discriminación.

   - Combinar varios bots y herramientas de Scraping con el fin de contrastar los resultados de cada uno.

   - Analizar la viabilidad de complementar el análisis con Web Crawling, con el fin de poder visitar varios enlaces a la vez, realizar análisis multinivel y estar actualizando el contenido almacenado por medio de rastreo.

2. Ampliar el entrenamiento de los modelos de aprendizaje supervisado y no supervisado con la combinación y comparación de técnicas, en ese sentido se sugiere:

   - En el modelo de Zero Shot Classifier, combinar etiquetas para los pasos 1 y 2, y revisar la funcionalidad de las etiquetas del segundo ejercicio "tengo algo que decir sobre el gobierno, los políticos escuchan lo que tengo que decir". Estas podrían ser más específicas y direccionadas.

   - En el análisis no supervisado, ampliar tanto la muestra de los comentarios como los motivos de discriminación. Previo a ello, se podría incluso hacer una preclasificación de dichos motivos de acuerdo con los resultados del análisis de Web Scraping.

   - Revisar las etiquetas del segundo ejercicio o explicar su funcionalidad en virtud de mapear lo que se ha logrado en el ODS 16.

3. Incluir en posteriores etapas del proyecto, el análisis de imagen y video. Ello teniendo en cuenta que gran parte del contenido que se encuentra hoy en la web se halla en estos formatos.

   - Una vez ejecutado el Web Scraping y Web Crawling, se puede crear una base de datos de imagen relacionadas con la discriminación o palabras cercanas, con el fin analizar de manera posterior, los datos mediante modelos de Deep Learning, preferible redes neuronales.

   - Para el entrenamiento del algoritmo en este caso, se hace necesario igualmente, una

parametrización de lo que se esperaría encontrar en las imágenes que se recopilen, en función de lo que nos dice la teoría sobre discriminación (desde lo sociológico y etnográfico) para luego si direccionar el modelo.

4. En una fase posterior del proyecto, se recomienda incluir el análisis de sentimientos y el análisis de discurso de odio.
   - Para lo anterior, las redes sociales que podrían arrojar mejor información son Twitter, Tik Tok y Facebook.
   - Los modelos más recomendados en este punto son: K-Vecinos Más Cercanos, Máquinas de Soporte Vectorial y Redes Neuronales.
   - Previo a la construcción del modelo y su respectivo entrenamiento se debe realizar un análisis desde las definiciones sociológicas o antropológicas sobre el discurso del odio relacionado con la discriminación.

Las siguientes recomendaciones están relacionadas con los resultados del proyecto:

1. La discriminación política debe ser parametrizada en intervalos de tiempo, dado que los resultados pueden estar afectados por la coyuntura política. En esta medida se recomienda revisar los siguientes aspectos:
   - Fragmentar los resultados por periodos de análisis y controlar por grupo poblacional.
   - Dado que es la categoría más representativa habría lugar a argumentarla desde lo sociológico.
   - Revisar si el porcentaje de comentarios de las categorías menos representadas se debe a la muestra seleccionada, o la red / fuente de datos escogida.
2. El análisis horizontal con modelo no supervisado Zero Shot, se podría realizar excluyendo la variable política y dando lugar a enfatizar otras categorías como la económica, creencias e identidad cultural. Frente a este aspecto se sugiere:
   - Fragmentar los resultados por periodos de análisis y controlar por grupo poblacional.
   - Verificar los resultados en función de las categorías o etiquetas establecidas en el modelo y el referente teórico en materia de discriminación.

3. La discriminación política debe ser parametrizada en intervalos de tiempo, dado que los resultados pueden estar afectados por la coyuntura política. Al respecto:

    - Se sugiere fragmentar los resultados por periodos de análisis y controlar por grupo poblacional.

    - Dado que es la categoría más representativa habría lugar a argumentarla desde lo sociológico.

4. Se sugiere estudiar la posibilidad de analizar el cruce entre categorías con el fin de poder establecer posibles asociaciones entre grupos o tendencias:

    - Para ello sería necesario crear etiquetas que combinen varias categorías de discriminación en el modelo de aprendizaje no supervisado.

    - Se sugiere fragmentar los resultados por periodos de análisis y controlar por grupo poblacional.

5. Revisar las categorías y etiquetas para las dimensiones de inclusividad, receptividad y representación política. Para ello se sugiere:

    - Consultar referentes teóricos que respalden las etiquetas de cada categoría dese la sociología y politología.

    - Para el tema de inclusividad, explorar otros métodos como el análisis de sentimientos y el discurso del odio, utilizando algunas técnicas o modelos mencionados con anterioridad.

Por último, vale la pena señalar los retos, lecciones aprendidas del proyecto y sugerencias generales, en particular, los siguientes aspectos:

- Se sugiere generar una ampliación de la muestra de datos y separación por grupos etario o poblacionales para enriquecer el análisis.

- Incursionar el análisis utilizando otras redes sociales y modelos de aprendizaje no supervisado, o la combinación de estos.

- Se sugiere combinar el uso de Web Scraping con Web Crawling con el fin de realizar análisis multinivel en diferentes páginas.

- En fases posteriores, se recomienda incluir en los análisis imagen y video, dado el gran contenido que ahora se publica en redes sociales en estos formatos.

- Como se mencionó en las sugerencias metodológicas, las etiquetas y categorías deben estar respaldadas por un referente teórico en el cual se sustenten, y puedan ser relacionadas o comparados con otros análisis.

- El análisis de fuentes secundarias debe fortalecer o ampliar la explicación que se da a través de las fuentes primarias y encuestas tradicionales, en consecuencia, se debe mantener marcos de referencia iguales o similares para que puedan ser complementarios y se logre tener una panorámica general de la problemática a analizar, en este caso el tema de discriminación.

Proyecto de "Índice de noticias como estimador temprano de la actividad económica"

Este proyecto tiene por objetivo probar un conjunto de modelos (Support Vector Regression, Lasso, Ridge, Random Forest Regressor, Red Neuronal) que permitan mejorar el cálculo de un estimador temprano para conocer el comportamiento aproximado del sector agrícola según las Divisiones CIIU Rev. 4 A.C.-61 agrupados utilizando modelos de machine learning y redes sociales.

En ese sentido, desde la Sala de Modernización Tecnológica se trabajó de manera conjunta con el equipo del DANE para escalar dicha metodología a los diferentes sectores de la economía colombiana con el ánimo de diseñar e implementar un índice de medición temprana del comportamiento económico que incorpore la información de fuentes de datos no convencionales como Google News y Google Trends para realizar el seguimiento a los diferentes sectores económicos. En un primer piloto se probó con el sector agrícola de Colombia y de ello se generó paper de investigación de autoría conjunta entre el DANE y la Universidad EAN.

El objetivo del paper "Colombian agricultural sector's early estimator of GDP post-pandemic COVID19 using Google News and Google Trends" es la documentación de los resultados del proyecto, en un artículo científico susceptible de ser publicado en el International Journal Forecasting.

Recomendaciones de los expertos al proyecto de "Índice de noticias como estimador temprano de la actividad económica"

Como recomendación general para proyectos enfocados en la automatización de procesos de calidad del lago de datos, se propone revisar la propuesta de valor que ofrece esa alternativa de automatización frente a lo que ya existe y definir una hipótesis de implementación del software para determinar el plan de trabajo del respectivo proyecto.

Por otro lado, desde la perspectiva del Big Data y la inteligencia artificial, se recomienda revisar los análisis de necesidades del usuario para implementar cualquier proceso tecnológico, teniendo en cuenta los requerimientos, el perfil de los usuarios y los usos que se le vayan a dar a los productos de los proyectos.

Asimismo, es importante evitar reprocesos puesto que, si constantemente se hacen análisis a los desarrollos de software, se pueden identificar errores en el camino, y en el resultado, el usuario quedará satisfecho sin hacer mayores cambios, lo cual es eficiente para el desarrollo de más y mejores proyectos.

Recomendaciones de los expertos al proyecto "Desarrollo y distribución de aplicaciones Web para la recolección de datos"

Desde la Sala de Modernización Tecnológica de la Producción Estadística del CASEN, se apoyó en su etapa preliminar, el proyecto de "Desarrollo y distribución de aplicaciones web para la recolección de datos" que tiene por objetivo desarrollar y distribuir, por medio de una red social, una aplicación web, como método alternativo de recolección de información estadística oficial. Su alcance se enfoca en la aplicación en versión beta de un instrumento y método de recolección que permita validar la viabilidad de escalar a otros productos de información del DANE; para ello, se desarrolló una prueba en un caso de recolección de datos para un indicador de Desarrollo Sostenible. De esta presentación, surgió la iniciativa de desarrollar una serie de seminarios web orientados a fortalecer el uso de metodologías ágiles para el análisis y validación del prototipo de la aplicación, así como en el tema de protección de datos y normatividad para la gestión de la información en la nube.

Interoperabilidad

Entre las recomendaciones brindadas por los expertos durante el año 2022 al Sistema Estadístico Nacional frente a la línea de Interoperabilidad, se destacan:

1. ***Generales***

   - Dar continuidad a los proyectos e iniciativas que se han venido implementando en materia de modernización tecnológica del SEN, en particular, estructurar e implementar protocolos para que las diferentes entidades adscritas al SEN logren avanzar y madurar en la interoperabilidad para la gestión de la información.

   - Se recomienda avanzar en el análisis de la línea de interoperabilidad desde una perspectiva multisectorial en la cual intervengan los diferentes actores del SEN, evaluando los avances de cada uno de cara a la implementación de la plataforma X-ROAD.

- Avanzar en la implementación del modelo de interoperabilidad, en la medida que articule y apalanque su desarrollo digital en tres pilares fundamentales: 1. Agilidad en los procesos de estandarización, normalización e intercambio de datos, 2. Eficiencia en el uso de los recursos e integración de la información, y 3. Intercambio de información protegiendo en todo momento los datos personales de los usuarios y ciudadanos adscritos al sistema.

- La puesta en marcha e implementación del Sistema Nacional de Interoperabilidad, a través de su correspondiente plataforma X-ROAD, bajo el liderazgo del Ministerio de Tecnologías de la Información y las Comunicaciones, implica tener en cuenta aspectos como el grado de conectividad de las instituciones y su nivel de madurez digital, la colaboración y grado de apertura para el trabajo interinstitucional, los procesos de gestión del cambio implementados en cada organismo y el respectivo marco de gobernanza que dará las pautas para dirigir el sistema (BID, 2019).

2. Frente al aspecto estratégico

- Generar espacios de diálogo y mapeo de las capacidades institucionales en relación con sus sistemas de información.

- Identificar el nivel de madurez de las entidades del SEN, respecto de la cultura organizacional basada en datos. Se subraya la importancia de trascender el nivel nacional y profundizar en los niveles regional y local.

3. Frente al dominio político – legal

- Se recomienda armonizar las prioridades y necesidades en el uso y gestión de la información en las entidades a los tres niveles. Ello mediante la programación de mesas de trabajo y diagnósticos colaborativos.

- Se sugiere, previo a la implementación del sistema y la plataforma de interoperabilidad, trabajar en la homogenización y estandarización de los sistemas de información a lo largo de la cadena de valor en la producción de datos (desde la captura a la toma de decisiones).

4. Frente al aspecto táctico

- Estructurar un programa de articulación entre las entidades del Estado que comprenda las siguientes fases: sensibilización, alistamiento, integración y escalamiento del modelo de interoperabilidad.

- Abrir espacios de diálogo y construcción colectiva en primera fase a nivel nacional, segunda fase a nivel regional o departamental y tercera fase a nivel local o municipal.

- Implementar un modelo de aprendizaje colaborativo en el uso de los Sistemas de Información para los diferentes actores y usuarios que se verán afectados o beneficiados con el sistema de interoperabilidad.

- Mapear las necesidades en el uso y gestión de la información a nivel regional y local.

- Desarrollar e implementar una metodología de evaluación de la calidad del dato para transferirla a los niveles regional y local.

- Desarrollar pilotos para la integración y uso de herramientas de Big Data para la gestión de los sistemas de información.

- Evaluar los resultados en la implementación de los pilotos y analizar la viabilidad para su escalabilidad.

- Implementar una estrategia continuada para la formación del ciudadano en la cultura de la información.

- Apropiar y adaptar metodologías para el uso escalonado en la apertura y tratamiento de datos personales.

4.2 Seguridad de la información

Uno de los grandes retos de la sociedad de la información es la salvaguarda de los datos como su principal recurso. En este contexto, las entidades públicas y las instituciones privadas deben hacer frente al reto de gestionar y garantizar la seguridad informática de los diferentes sistemas de información y bases de datos que manejan. No solo porque la información y los datos se han convertido en el principal activo de las organizaciones, sino también por los derechos de privacidad y habeas data relativos a la información personal, de los cuales son sujetos todos los ciudadanos.

En esta medida, en la Sala de Modernización Tecnológica se discutió sobre los retos y desafíos a los cuales se enfrenta el Sistema Estadístico Nacional, en materia de seguridad informática; haciendo énfasis tanto en la importancia de tener la tecnología adecuada para responder de manera eficaz ante las diferentes amenazas que podrían vulnerar la salvaguarda de la información personal de los usuarios, como en el marco normativo y las leyes que protegen a los individuos en relación a sus datos personales.

En el contexto de las sesiones de la sala durante el 2022, el Departamento Administrativo Nacional de Estadística -DANE, en cabeza de los pares de la Oficina de Sistemas, presentó ante los expertos, el contexto de la entidad y las correspondientes inquietudes frente a la gestión de la información en la nube y seguridad de los datos, así como los requerimientos de apoyo en materia jurídica para tal área.

En desarrollo de la línea de investigación de seguridad de la información, se realizaron cinco (5) reuniones en donde se abordaron los siguientes temas:

1. Presentación del contexto del DANE en materia de seguridad de datos gestionados en la nube y protección de datos confidenciales, a fin de recibir recomendaciones por parte de los expertos y fortalecer estos procesos (13 de mayo 2022).

2. Reunión sobre el aspecto normativo de la seguridad de los datos con expertos externos en el tema (17 de junio 2022).

3. Reunión sobre el aspecto técnico de la seguridad de los datos con el grupo de cyberseguridad de la Universidad de los Andes (17 de junio 2022).

4. Reunión interna para establecer una interlocución entre la Oficina de Sistemas del DANE y la experta Andrea Martínez (invitada a la sala) para resolver dudas frente a las restricciones de negocio para la seguridad de los datos al adoptar nube (14 de julio de 2022).

5. Oportunidades de uso de blockchain en operaciones estadísticas y la presentación del marco ético de los datos (22 de noviembre 2022).

Al respecto, surgieron las siguientes recomendaciones:

1. Recomendaciones generales

- En materia de seguridad de los datos, surge la oportunidad de implementar un plan de

ciberseguridad a nivel de las diferentes instancias del SEN, que involucre tanto el componente técnico como el jurídico para el manejo y tratamiento de los datos personales de los usuarios.

- Se recomienda que, en la nueva vigencia, se genere una estrategia integral para la seguridad de la información y la protección de los datos en la nube de los ciudadanos, la cual contenga tanto los protocolos, como el marco normativo y su manual de operación actualizado a los diferentes retos que hoy presenta la ciberdelincuencia y las amenazas externas.

2. Recomendaciones sobre el aspecto técnico de la seguridad de los datos en nube

- Se sugiere revisar la tendencia actual de "privacidad diferencial" para determinar el modelo adecuado para el procesamiento de datos en nube de acuerdo con las características de la entidad.

- Desde un abordaje conceptual, se recomienda separar conceptualmente la "seguridad" y la "privacidad" de los datos.

- Se sugiere que, previo a la migración de los datos a la nube, se verifiquen las necesidades, los requerimientos y se prevean las consecuencias. Es decir, es necesario discutir los beneficios, debido a la mejora en la capacidad de cómputo en nube y los riesgos que genera.

3. Recomendaciones sobre el aspecto normativo de la seguridad de los datos en nube

- Se debe tener en cuenta que los principios de reserva estadística aplican aún para las excepciones.

- Al momento de migrar datos a la nube, el criterio básico de seguridad es tratar todo como si fuera información personal, es decir, aplicar el máximo nivel de seguridad.

- Desde la experiencia, se evidencia que las entidades públicas pueden usar la nube y subir cualquier tipo de información sin ninguna restricción legal.

- Se recomienda al DANE, blindar los contratos con los proveedores de nube (debida diligencia) con anexos robustos que garanticen que el proveedor cumpla con la normatividad local.

- Se sugiere revisar casos de buenas prácticas en materia de gestión de datos en nube en el sector público, entre los que se destacan: Dirección de Impuestos y Aduanas Nacionales -DIAN, Ministerio de Salud y Protección Social, Colombia Compra Eficiente y Secretaría de Educación de Bogotá.

Bibliografía

Bureau de Estadísticas Australiano, (2021). *Privacy, Confidentiality & Security*. Recuperado de https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Privacy,+Confidentiality+&+Security

Departamento Administrativo Nacional de Estadística, DANE. (2012). *Adaptado de Clasificación Industrial Internacional Uniforme de todas las actividades económicas CIIU Rev 4. A.C.* Bogotá. Recuperado de https://www.dane.gov.co/files/sen/nomenclatura/ciiu/CIIU_Rev4ac.pdf

Departamento Administrativo Nacional de Estadística, DANE. (2012). Clasificación Internacional Uniforme de todas las actividades Económicas CIIU Rev.4.AC. Bogotá. Recuperado de https://www.dane.gov.co/files/sen/nomenclatura/ciiu/CIIU_Rev4ac.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones, MinTic. (2019). Marco de Interoperabilidad para gobierno digital. Recuperado de https://www.mintic.gov.co/arquitecturati/630/articles-9375_recurso_4.pdf

Organización para la Cooperación y el Desarrollo Económicos, O. (2000). *Organización para la Cooperación y el Desarrollo Económicos OCDE*. Recuperado de Basado en la definición de Censo Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva: https://stats.oecd.org/glossary/detail.asp?ID=301

Pombo, C. Ortega, G. Olmedo, F. Solaine, M. Cubo, A. (2019). El ABC de la interoperabilidad de los servicios sociales marco conceptual y metodológico. Banco Interamericano de Desarrollo, Washington D.C, 2019