



DANE

SEN^{2.0}
Sistema Estadístico
Nacional-Colombia

CONSEJO TÉCNICO ASESOR DEL SISTEMA ESTADÍSTICO NACIONAL - CASEN

SALA ESPECIALIZADA PARA LA MODERNIZACIÓN TECNOLÓGICA DE LA PRODUCCIÓN ESTADÍSTICA

DOCUMENTO DE RECOMEDACIONES Y PRODUCTOS GENERADOS POR LA SALA

Diciembre de 2024

**DEPARTAMENTO ADMINISTRATIVO
NACIONAL DE ESTADÍSTICA
-DANE**

B. PIEDAD URDINOLA CONTRERAS
Directora

ANDREA RAMÍREZ PISCO
Subdirectora

ÁLVARO FERNANDO GUZMÁN LUCERO
Secretaria General

DIRECTORES TÉCNICOS

CÉSAR MAURICIO LÓPEZ ALFONSO
Dirección de Metodología y
Producción Estadística

JAVIER SEBASTIÁN RUIZ SANTACRUZ
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

ELKIN ERNESTO RAMÍREZ NIÑO
Dirección de Geoestadística

DIANA MARÍA BOHÓRQUEZ LOSADA
Dirección de Difusión, Mercadeo y
Cultura Estadística

© DANE, 2024

Prohibida la reproducción total o parcial sin
permiso o autorización del Departamento
Administrativo Nacional de Estadística
Colombia

MARIO LINARES VÁSQUEZ
NICOLÁS CARDOZO ÁLVAREZ
LEÓN DARIO PARRA BERNAL

Expertos

Consejo Técnico Asesor del Sistema
Estadístico Nacional -CASEN

MÓNICA PATRICIA PINZÓN TORRES
DERLY VIVIAN LIZARAZO GARCÍA
Coordinación de Planificación y Articulación
Estadística

CONTENIDO

Introducción4

1. OBJETIVO GENERAL.....6

2. ANTECEDENTES6

3. RECOMENDACIONES.....8

3.1 Ciencia de datos – Ciclo de Vida8

3.2 Ciencia de Datos – Proyectos de Ciencia de Datos..... 12

4. SEGURIDAD Y CUMPLIMIENTO 15

4.1 Anonimización de Datos..... 15

4.2 Interoperabilidad 16

5. PRODUCTOS DE LA SALA 19

INTRODUCCIÓN

El Departamento Administrativo Nacional de Estadística -DANE, a partir de la Ley 2335 de 2013 en su artículo 14, establece la creación del Consejo Asesor Técnico del Sistema Estadístico Nacional (CASEN), que tiene como objeto asesorar al DANE y al Gobierno Nacional sobre cuestiones de importancia estratégica para las estadísticas oficiales del país. Está compuesto por la sala general y cinco salas especializadas en diferentes temáticas: 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.

La sala especializada de Modernización tecnológica de la producción estadística está conformada por los expertos Mario Linares Vásquez, Nicolás Cardozo Álvarez y León Darío Parra Bernal, Julieth Alejandra Solano Villa como secretaria técnica de la sala y los delegados de las diferentes direcciones del DANE: Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN), Dirección de Síntesis y Cuentas Nacionales (DSCN), Dirección de Metodología y Producción Estadística (DIMPE) y Oficina de Sistemas (OSIS). La metodología que se utilizó para el desarrollo de la sala consistió en reuniones plenarias presenciales y virtuales, en donde se presentaron los avances en temas de interés en el contexto de la entidad y se propusieron preguntas orientadoras, con el objetivo de recibir realimentación por parte de los expertos.

En línea con lo anterior, con apoyo de la sala se realizaron dos seminarios, el primero “Arquitectura de datos: fundamentos y buenas prácticas” y el segundo seminario “Guía para la anonimización de datos estructurados”, como complemento de las líneas de investigación.

Este documento tiene como objetivo presentar al Sistema Estadístico Nacional (SEN) y a la sala general del CASEN, las recomendaciones realizadas por los miembros de la Sala Especializada de Modernización Tecnológica de la Producción Estadística y el DANE, frente a las líneas de investigación propuestas: i) Ciencia de datos – Ciclo de vida, ii) Ciencia de datos – Proyectos de datos; y iii) Seguridad y cumplimiento.

A continuación, se presentan el objetivo general del documento, los antecedentes y concluye con las recomendaciones de los expertos para las líneas de investigación trabajadas y los productos generados por los miembros de la sala en el 2024.

OBJETIVO GENERAL

Presentar a la sala general del CASEN, las recomendaciones realizadas por los miembros de la Sala Especializada de Modernización Tecnológica de la Producción Estadística durante el 2024, en las líneas de investigación: i) Ciencia de datos – Ciclo de vida, ii) Ciencia de datos – Proyectos de datos; y iii) Seguridad y cumplimiento.

ANTECEDENTES

Para la propuesta del plan de trabajo del 2024 se tuvo en cuenta los temas tratados los anteriores años del CASEN como lo son interoperabilidad y seguridad de la información, haciendo énfasis en la elaboración de productos que brinde herramientas al SEN, que permitan emplear técnicas y métodos de análisis avanzados para mejorar la recopilación, el procesamiento y la presentación de datos estadísticos producidos desde el DANE; así como garantizar de manera oportuna la integridad, la privacidad y la confiabilidad de los datos estadísticos generados por el DANE.

En este sentido, en la sala de modernización se definió trabajar las siguientes líneas investigación: i) Ciencia de datos – Ciclo de vida, ii) Ciencia de datos – Proyectos de datos; y iii) Seguridad y cumplimiento.

De acuerdo con el Banco Interamericano de Desarrollo (BID) la ciencia de datos, “... es el campo interdisciplinario que estudia los datos dentro de un área de conocimiento específica a partir de técnicas informáticas, estadísticas, matemáticas, analíticas y de visualización para extraer conocimientos que respalden la toma de decisiones informadas. En este proceso, se utilizan diversas herramientas, tecnologías y lenguajes de programación para el análisis de los datos y por lo general, la extracción y procesamiento se automatizan para facilitar su actualización o reproducción. Estos datos pueden abarcar una variedad de formas, como números, texto, imágenes o sonidos. La esencia de la ciencia de datos se centra en proporcionar un análisis y una presentación efectiva de los datos con el propósito de abordar desafíos concretos en el mundo real.”

Dentro de este contexto, la línea de Ciencias de Datos se dividió en dos temas:

1. Ciclo de vida de modelos de proyectos de ciencia de datos (temas transversales de inteligencia artificial y aprendizaje de máquina segura), que tendrá como resultado la Guía para la gestión y mantenimiento del ciclo de vida de modelos de proyectos de ciencia de datos
2. Proyectos de ciencia de datos: línea en la cual se continuó la revisión de proyectos trabajados en la pasada vigencia del CASEN: el índice de noticias, que tiene como objetivo generar un índice de contraste de las actividades económicas, a partir de fuentes no tradicionales. De dicho proyecto, se publicó el artículo titulado: Colombian Agricultural Sector's Early Estimator of Gross Domestic Production Using Nowcasting and Big Data Methods | Journal of Technology Management & Innovation en agosto de 2024. Igualmente, se avanzó en la y la App diversa cuyo objetivo es medir la probabilidad de que la gente experimente eventos de discriminación a partir de los datos de la Encuesta de Cultura Política y del Censo Nacional de Población y Vivienda 2018, como el tema anterior se pretende robustecer con las recomendaciones dadas por la sala.
3. La tercera línea de investigación, seguridad y cumplimiento se enfoca en garantizar de manera oportuna la integridad, privacidad y confiabilidad de los datos estadísticos generados por el DANE. En el desarrollo de esta línea, se trabajarán dos temas: i). Anonimización de datos, al respecto se menciona que a partir de la guía de Anonimización elaborada y actualizada por DIRPEN, se quiso solicitar recomendaciones para la inclusión de técnicas innovadoras y emergentes de Anonimización con enfoque tecnológico y fortalecer el documento para ser socializado con las entidades del SEN, y ii). interoperabilidad, en donde se quiere como producto la Guía para el proceso de interoperabilidad, un documento con mejores prácticas para el uso de plataformas en los procesos de interoperabilidad. Definidos los temas, se iniciaron las reuniones de trabajo de la sala.

RECOMENDACIONES

3.1 Ciencia de datos – Ciclo de Vida

El objetivo de esta línea de investigación está centrado en emplear técnicas y métodos de análisis avanzados para mejorar la recopilación, el procesamiento y la presentación de datos estadísticos producidos desde el DANE.

En el marco del trabajo liderado por la Oficina de Sistemas del DANE y con el apoyo del CASEN, se elaboró la “Guía para la gestión y el mantenimiento del ciclo de vida de modelos de proyectos de datos”, publicada en la página del DANE en el enlace <https://www.dane.gov.co/files/sen/lineamientos/guia-gestion-mantenimiento-ciclo-vida-modelos-20-proyectos-ciencia-datos.pdf> y en la página del SEN <https://www.sen.gov.co/sites/default/files/pagina-migraciones-files/2024-11/guia-para-la-gestion-y-el-mantenimiento-del-ciclo-de-vida-de-modelos-de-proyectos-de-ciencia-de-datos.pdf>. En este trabajo se realizaron 5 reuniones en las cuales la OSIS presentaba los avances del documento a los miembros de la sala para su realimentación.

A continuación, se presenta las recomendaciones realizadas por los expertos:

Definición del marco teórico asociado

En esta sección, se realizó la revisión y discusión de conceptos claves en el ciclo de vida de modelos de proyectos de ciencia de datos, identificando las mejores prácticas y tendencias actuales en la industria, detallando las fases del ciclo de vida de un proyecto de ciencia de datos y la importancia y desafíos específicos de estas.

Recomendaciones:

- Incluir los conceptos de: ciclo de vida: inteligencia artificial generativa; inteligencia artificial general; modelos de lenguaje de gran tamaño (LLMs), inteligencia artificial responsable, modelo, aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo y Data cleaning.
- No considerar las metodologías ágiles por ser lineales y burocráticas, como: Knowledge Discovery in Databases (KDD) y Cross-Industry Standard Process for Data Mining (CRISP-DM), en lugar a Team Data Science Process (TDSP): sugirieron revisar las metodologías

AN Analytics, Guerrilla Analytics y tener en cuenta autores como Scott Ambler y Martin Fowler.

- Tener en cuenta el marco de trabajo Scrum, muy utilizada en proyectos tecnológicos que se pueden combinar con otras metodologías como la de pensamiento – diseño y hacer match con las mencionadas anteriormente; resaltando la importancia de combinar lo técnico y la gestión de proyectos, para dar agilidad en su implementación.
- Realizar un proceso de versionamiento para reconocer los avances y las diferencias presentadas, frente a esto los expertos compartieron un artículo sobre cómo aplicar la ingeniería de software para machine learning, con el fin de evaluar si se puede alinear con el TDSP.
- Contar con una metodología para evaluar la calidad de los datos y generar una limpieza y estandarización de estos, con el fin de mejorar los modelos.
- Precisar en la guía, si los datos van a un proceso de fusión, se requiere incluir la programación, la implementación y el diseño de las fuentes. Asimismo, para evaluar modelos se deben enfocar en cuatro atributos: ¿qué tan bueno es el modelo para la tarea?; el performance visto desde el rendimiento a consumir recursos; la seguridad, y el lidiar con los datos antagonistas.
- Usar bandax combinada con phyton para la limpieza de los datos, ya que puede hacer más amigable la usabilidad de los modelos y ayuda con la calidad de los datos. Al respecto, se sugirió conocer la experiencia de los ingenieros de la iniciativa de Augusta.
- Contar con esquemas frente al manejo del proceso de incidencias, cómo se llevan a cabo las actividades de revisión, solución y soporte a estas incidencias, con el fin de realizar ejercicios post mortem que permitan hacer una gestión más efectiva del conocimiento.
- Incluir una línea de trazabilidad en el ciclo de vida de cada proyecto de ciencia de datos sobre análisis y seguimiento al usuario final, con el fin de probar las diferentes versiones del proyecto en relación a las necesidades de los diferentes usuarios o actores clave.

Exploración de herramientas y plataformas para cada etapa del ciclo de vida de los modelos de ciencia de datos

En este espacio se presentó un resumen de los aspectos claves, beneficios y limitaciones de las principales herramientas y plataformas utilizadas para gestionar cada etapa del ciclo de vida de modelos de ciencia de datos, se recibió la siguiente realimentación por parte de la sala:

Recomendaciones:

- Utilizar en la gestión de paquetes y entornos virtuales: Conda, Python Package Index (PyIP), Python con Pyenv y Poetry, Renv y Pactrak los procesos de verificación (Safety) y hacerlos parte de Jupiter o el Colab.
- Tener en cuenta en el control de versiones de código, diferenciar lo que es el sistema de control de versiones y qué son las plataformas para control de versiones, incluir el de Microsoft (Azure), cambiar lo que se hacía con SVN por GIT y usar el sitio web Github, ya que cuenta con sus propios protocolos para la fuga de información.
- Revisar Kit LFSY, Dolto, DLT o GIT FNS, para el control de versiones de datos.
- Incluir Datadog, un servicio de monitorización para aplicaciones en la nube, en el desarrollo, el monitoreo y la gestión de los proyectos.

Revisión de prácticas de Ingeniería de Software para Machine Learning

En esta sección se presentaron los aspectos claves de las metodologías de ingeniería de software para Machine Learning utilizadas en el ámbito académico y empresarial que abordan todas las etapas del ciclo de vida de modelos de ciencia de datos. Asimismo, se detallaron las etapas y las mejores prácticas de desarrollo de código que conforman las metodologías de ingeniería de software para Machine Learning.

A continuación, se relacionan las recomendaciones de los miembros de la sala sobre las prácticas de ingeniería de software para Machine Learning.

Recomendaciones:

- Incluir en la participación por roles, un rol en planeación, como un ingeniero de datos o un científico de datos y un analista, con el fin de que puedan opinar sobre lo programado y evitar reprocesos posteriores. Asimismo, un ingeniero de desarrollo o de software, que hará parte del equipo que se dedicará a desplegar, embeber el modelo dentro de un sistema más grande o en su defecto desarrollar una interfaz más adecuada e incluir a una persona dedicada a la responsabilidad, haciendo referencia a la ética, la robustez y la seguridad, ya que existen certificaciones para validadores de responsabilidad de implementación de ML o inteligencia artificial.
- Incluir en la validación temprana la participación de los stakeholders, ya que ellos son los que aceptan el trabajo que se está haciendo.

- Definir cuál será el estándar de las interfaces del modelo: ¿cómo se va a llamar el modelo? ¿qué va a retornar? ¿cuáles van a ser las salidas?, a partir de esto el código de los modelos contará con una clase para todos los modelos o los proyectos en la API.
- Publicar el modelo como una librería, que contenga una función con una definición clara de las entradas y las salidas.
- Documentar los modelos o dentro del código, manejar un conjunto de documentación externa que puede ser algo simple como una wiki en el repositorio donde manejan el código o que se defina un estándar de cómo se puede documentar.

En relación con la estructura del árbol de codificación:

- Tener en cuenta en el árbol la visualización y un notebook adicional que indique el orden en que se debe proceder con lo respectivo. Asimismo, se sugirió agregar un marco de principios y valores a la guía, y una metodología como Extreme Programming, que da lineamientos de cómo organizar el código; la guía debe tener la suficiente versatilidad para ajustarse dependiendo de las necesidades del equipo interno y la visión del usuario.
- Dejar documentado todo el proceso de experimentación del modelo, antes de llegar a los parámetros adecuados (llevar una bitácora de las pruebas realizadas).
- Construir un Massive Open Online Course (MOOC) que en una primera versión tenga vídeo, animación o un contenido estático que presente la consolidación de herramientas que se pueden usar.

Especificación de los artefactos

En este aparte se presentaron los artefactos definidos en el ámbito académico y empresarial que respaldan la documentación de un proyecto de ciencia de datos, detallando el contenido de la documentación asociada a los procesos que integran las etapas del ciclo de vida de un modelo de ciencia de datos. A su vez, se discutió la propuesta de documentación que mejor se acople a las necesidades de la Entidad respecto a la gestión e implementación de proyectos de ciencia de datos.

Recomendaciones:

- Documentar las pruebas con el conjunto de datos utilizado y la ubicación, entre otras; es importante que quede un registro de las pruebas realizadas, para dar a conocer con qué conjuntos de datos se realizaron y la replicabilidad del ejercicio.
- Incluir en el documento marco, la infraestructura teniendo en cuenta las expectativas para un buen término, los objetivos, en donde estamos y qué necesitamos, las características del despliegue incluyendo presupuestos, este puede ser un ejercicio a cargo del Project Manager, que sirva de referencia a la infraestructura ideal.
- Incluir en el informe resumen de datos, consideraciones de seguridad, éticas y de tamaño, dependiendo el tipo de información, si se hace necesario anonimizar, si son datos sensibles (disclaimer). Asimismo, las características de los metadatos.
- Colocar la justificación del porqué del modelo utilizado; también tener por separado las métricas de evaluación, la validación cruzada, las pruebas de robustez, el análisis de sesgo, varianza y las pruebas de interpretabilidad del modelo.
- En el informe resumen de despliegue, no es conveniente utilizar en el contenido “consideraciones para el mantenimiento del modelo”, es mejor eliminar la palabra modelo y colocar del despliegue, esto permite que sea entendible para las personas de operaciones o técnicos y detallar como sería la liberación del modelo. Por otra parte, este item se podría desglosar: escalabilidad del modelo, mejora continua, gestión de incidentes, reentrenamiento y consideraciones del monitoreo post del modelo.

3.2 Ciencia de Datos – Proyectos de Ciencia de Datos

Esta línea se centra en emplear técnicas y métodos de análisis avanzados para mejorar la recopilación, el procesamiento y la presentación de datos estadísticos producidos desde el DANE

Índice de Noticias

En la presentación del tema se resaltó el uso de fuentes alternativas de información a través del índice de noticias de las actividades económica, se esperaba contar con recomendaciones para robustecer y precisar el modelo del índice de noticias que permita replicarlo.

Recomendaciones:

- Usar técnicas no supervisadas para la clasificación de las noticias en lenguaje natural. Otra opción es usar un modelo semántico donde periódicamente se pueden ir refinando los términos de la búsqueda, esta opción busca subir la calidad.

- Hacer un cluster donde el sentimiento general sea positivo o negativo, para no quedarse con una fuente en particular, no usar solo el concepto, el título; sino también el texto y buscar similitud entre las noticias.
- Hacer un análisis cualitativo antes de correr un modelo, con una muestra de manera reiterada y después mejorar la parte semántica. Además, es fundamental revisar la calidad del dato antes de aplicar cualquier modelo.
- Incluir entre el procesamiento y el análisis un paso donde se revise la noticia o el tópico con un filtrado automático y se analice dentro del conjunto de noticias a fondo en cada uno de esos grupos.
- Identificar la fuente más fidedigna (noticias, twitter, blogs) a utilizar dependiendo del tipo de análisis que se va a realizar y la temporalidad.
- Tomar los titulares, agrupar los que mencionan el mismo tema y por volumen se puede o no controlar un sesgo en particular, dependiendo los porcentajes; esto permite mirar lo positivo y lo negativo y tomar la decisión dependiendo de eso.
- Tener en cuenta todas las líneas exploratorias y el análisis cualitativo, respetando la línea del proyecto y el tiempo.
- Explorar nuevas técnicas de análisis de sentimientos para encontrar interrelaciones entre los términos de un tema determinado, con el fin de enriquecer la información que se le entrega a los analistas de coyuntura o tomadores de decisiones para los cuales es diseñado el índice.

App Diversa

Herramienta construida por el DANE para medir la probabilidad que tiene alguna persona de experimentar discriminación. La App Diversa, es un proyecto que busca comprender el fenómeno social, y que pretende alcanzar uno de los Objetivos de Desarrollo Sostenible, contemplados en la agenda 2030 de la Organización de las Naciones Unidas.

El piloto de la aplicación se presentó como una innovación en los métodos de recolección estadística, generando una visibilidad de esta problemática, y que supone, creará una toma de conciencia acerca de dicha problemática social.

Recomendaciones:

- Herramientas de desarrollo: utilizar Power Apps de Microsoft como una alternativa complementaria a Drupal, dada su integración con el ecosistema de Microsoft y su facilidad de uso.
- Alcance: ampliar el alcance de la aplicación para recopilar información de diferentes maneras innovadoras, no solo limitándose al tema de la discriminación. Se mencionan ejemplos como la percepción de la calidad de los servicios públicos o prácticas relacionadas con el uso del agua en las comunidades.
- Diseño de la interfaz y proceso de desarrollo: estandarizar el diseño de la interfaz gráfica, la experiencia del usuario y el proceso de desarrollo en todos los sistemas o herramientas de recolección de información. Esto implica mantener una línea gráfica y de usabilidad consistente en todas las aplicaciones, así como definir una metodología de desarrollo clara y pruebas de calidad continuas.
- Experiencia del usuario: realizar pruebas exhaustivas para garantizar una experiencia de usuario fluida y evitar abandonos. Señalan que el diseño actual de la aplicación, aunque visualmente atractivo, puede presentar problemas de navegación y pérdida de información si el usuario se equivoca. Se recomienda implementar estrategias de "ganchos" para mantener el interés del usuario y facilitar la recuperación de información en caso de errores
- Estandarización: crear un manual de diseño o un libro de conocimiento que categorice las herramientas según el tipo de encuesta, estilo y atributos a considerar. Este manual serviría como guía para elegir la herramienta adecuada y garantizar la consistencia en el desarrollo de futuras aplicaciones
- Seguridad: prestar atención a la seguridad en herramientas de código abierto como Drupal, realizando pruebas continuas y manteniendo las actualizaciones al día. Se destaca la necesidad de verificar vulnerabilidades y aplicar actualizaciones para proteger la integridad de los datos recopilados.
- Incentivos: considerar mecanismos de enganche ("nudges") o incentivos para motivar la participación voluntaria en las encuestas. Se mencionan ejemplos como recompensas monetarias, emocionales o la posibilidad de recibir retroalimentación personalizada basada en los datos proporcionados
- Pruebas de aceptación: realizar pruebas de aceptación en diferentes navegadores y dispositivos para garantizar la compatibilidad y usabilidad de la aplicación. Se hace hincapié en la importancia de probar la aplicación en contextos reales de uso para identificar y solucionar posibles problemas.
- Guía de accesibilidad: basarse en una guía de accesibilidad para garantizar que la aplicación sea inclusiva y accesible para todos los usuarios, independientemente de sus

capacidades o discapacidades. Esto incluye aspectos como el diseño visual, la navegación y la compatibilidad con tecnologías de asistencia.

SEGURIDAD Y CUMPLIMIENTO

4.1 Anonimización de Datos

En este espacio se expusieron los conceptos, el proceso de Anonimización de bases datos del Sistema Estadístico Nacional (SEN), teniendo en cuenta los pasos previos como conformar el equipo de trabajo y los requerimientos inicial, se explicaron cada una de las seis fases del proceso de anonimización: etapa 1. Revisiones previas, etapa 2. Análisis de riesgo, etapa 3. Selección de técnicas, etapa 4. Análisis de viabilidad, etapa 5. Aplicación de técnicas y etapa 6. Evaluación de resultados y las actualizaciones realizadas a la guía de anonimización, que contemplaba el contexto, la motivación para la actualización de la guía, las necesidades y desafíos actuales, como se realizó el proceso de actualización y los principales cambios.

La realimentación recibida frente a los expuesto fue la siguiente:

Recomendaciones:

- Para el proceso de anonimización se pueden tener en cuenta sistemas apoyados por técnicas de aprendizaje en máquina, que permitan identificar las relaciones entre variables.
- Tener en cuenta para la anonimización de cifrados a partir de técnicas adversarias, aplicados a multimedia por ejemplo el modelo Segment Anything Model (SAM) de Meta, útil para otro tipo de información diferente a la tabular
- Diferenciar metodologías, técnicas y casos y con el equipo de ciencia de datos revisar una IA que este customizada a las necesidades y hacer pilotos para no dañar toda la base de datos.
- Incluir en la guía los perfiles, para en función de eso saber que privilegios se tendrán para conocer la información.
- Considerar el uso de la red generativa antagónica (GAN) que es una arquitectura de aprendizaje profundo. Entrena dos redes neuronales de modo que compitan entre sí para generar nuevos datos más auténticos a partir de un conjunto de datos de entrenamiento.

- Colocar en la guía las implicaciones que tiene cada una de las técnicas, colocar las debilidades y fortalezas, para que el usuario de acuerdo con sus necesidades pueda seleccionar la técnica indicada.

4.2 Interoperabilidad

El desarrollo de este tema contempló varios momentos, los cuales se presentan a continuación:

Marco teórico sobre procesos de interoperabilidad

Se presentaron los conceptos y su aplicación en la interoperabilidad, definición de requisitos y objetivos específicos para la implementación en los procesos de interoperabilidad.

Recomendaciones:

- Tener un equilibrio entre lo que se va a interoperar y como mantener la replicabilidad de la información. Así como, mencionar un punto sobre el buen uso de las herramientas y las recomendaciones de cuando se deben usar y cuando no.
- Poner en práctica las recomendaciones hechas por la sala en el año 2021-2022, donde se habían establecido líneas de acción y se habían propuestos proyectos piloto con la encuesta de Gasto de Turismo Interno (GTI), para identificar cómo funcionaba la Interoperabilidad. Así mismo, temas como economía naranja, indicadores SENA, con geoportales y SIPSA para el sector agropecuario.
- Hacer extensivas las recomendaciones al Ministerio de las Tecnologías de la Información y Comunicaciones (MinTIC), teniendo en cuenta lo propuesto en interoperabilidad del libro blanco de la Comisión Económica para América Latina y el Caribe (CEPAL) y Naciones Unidas que debían ser trabajadas en mesa técnicas enfocado a la implementación del X-Road.
- Preparar mesas técnicas con las demás salas especializadas del CASEN para la difusión del X-Road.
- Implementar mesas técnicas con diferentes actores del SEN con el fin de mapear su nivel de desarrollo y maduración en interoperabilidad, así como las buenas prácticas en la gestión y seguridad de la información.

Evaluar las plataformas de interoperabilidad existentes

Este tema contempla la presentación de la interoperabilidad realizada entre diferentes geoportales, por ejemplo, Interoperabilidad SDMX – Geoportales y Geovisores SIPSA – EGIT, ejemplos de proyectos de interoperabilidad activos en el DANE; como el de los Objetivos de Desarrollo Sostenible – ODS; la evaluación de las plataformas existentes de interoperabilidad como el X-road: los mecanismos de interoperabilidad e intercambio de Información DANE.

Recomendaciones:

- Recopilar, por ejemplo, en un framework las prácticas de integración, así como presentar los pasos a seguir.

- Diseñar algunos indicadores por cada solución tecnológica que permitan establecer, en la medida que algunas opciones son más convenientes que otras. De forma tal que se vayan generando también capacidades y alternativas para las entidades que puedan estar involucradas.
- Documentar los procedimientos teniendo en cuenta las bondades y debilidades de cada uno de los protocolos de publicación/llamado de los servicios (SOAP WS, RESTapi, SDMX-WS), los métodos, costos y beneficios.
- Solicitar al Ministerio de las Tecnologías de la Información y Comunicaciones (MinTIC), la publicación de la guía una vez esté finalizada.

Evaluación de riesgos

En esta sección se trató el tema de capacidad de diferentes sistemas, organizaciones y tecnologías para trabajar juntos de manera eficiente, intercambiando y utilizando información. Se presentaron los dominios de interoperabilidad a tener en cuenta: legal, organizacional, semántico y técnico y se recibieron comentarios para tener en cuenta en la elaboración de la guía:

Dominio de interoperabilidad

Enfocado en los marcos regulatorios, normativos y legales que respaldan o limitan el intercambio de datos entre sistemas.

Recomendaciones:

- Incluir en la guía un ejemplo de lo que puede suceder cuando se materializa un riesgo y las implicaciones que tiene dentro del proceso. En esta misma línea, contemplar en el futuro el manejo de contratos inteligentes.
- Tener en cuenta en la construcción de la guía, elementos como la normatividad, los procedimientos, las buenas prácticas que han manejado las diferentes instituciones, los sistemas de información y la gestión de incidentes.

Dominio organizacional

Abarca los procesos, estructuras y políticas que facilitan la interoperabilidad entre instituciones, teniendo en cuenta los riesgos y las estrategias de mitigación.

Recomendaciones:

- Proponer en la guía un marco para evaluar el nivel de madurez de cara a la interoperabilidad, que le permita a cada entidad identificar como se encuentra (herramienta de autodiagnóstico).
- Sugerir en la guía a las instituciones monitorear los riesgos organizacionales (semestral o anual) para ir observando la evolución de estos, como estrategia de mitigación

Dominio Técnico

Se enfoca en la infraestructura tecnológica, protocolos y estándares que permiten el intercambio de datos entre sistemas, teniendo en cuenta riesgos y estrategias de mitigación.

Recomendaciones:

- Agregar un riesgo adicional al de obsolescencia tecnológica, relacionado con que no se cuente con los recursos tecnológicos en el escenario.
- Aclarar en la guía en quien tenga a cargo la custodia de los datos para el proceso de interoperabilidad cuente con todas las medidas de seguridad necesarias.

5. PRODUCTOS DE LA SALA

Las recomendaciones dadas por los expertos de la sala se vieron reflejadas en la Guía para la gestión y el mantenimiento del ciclo de vida de modelos de proyectos de datos y en la Guía para el proceso de interoperabilidad. Así mismo en las mejoras del Índice de noticias y la App diversa.

Como productos complemento de las líneas de investigación, se realizaron dos seminarios el primero "Arquitectura de datos: fundamentos y buenas prácticas" se realizó el 20 de junio y su objetivo fue compartir con los actores del ecosistema de datos del Sistema Estadístico Nacional (SEN) y con los administradores de datos sectoriales del Estado, elementos de la arquitectura de datos, su importancia y las mejores prácticas para su implementación efectiva en el ciclo de

vida de los proyectos de la ciencia de datos, de los proyectos de aprovechamiento de datos y de su articulación para la gestión pública. En este espacio se contó con la participación de expertos como Jairo Alberto Riascos Muñoz, Harvey Rosas, Álvaro Mauricio Montenegro Díaz, Tito Pablo Neira Ávila y Vivian Lucia Aranda Camacho.

El segundo seminario “Guía para la Anonimización de datos estructurados”, realizado el 23 de octubre tuvo como objetivo presentar la actualización de la guía en el marco del Plan Nacional de Infraestructura de Datos (PNID) en trabajo conjunto del DANE y del Archivo General de la Nación (AGN) y contó con dos talleres prácticos realizados en la Universidad de los Andes.