

Consejo Asesor Técnico del Sistema Estadístico Nacional (CASEN) 2023-2025
Sala Especializada para la Modernización Tecnológica de la Producción Estadística
Acta No. 5

Ciudad: Bogotá D.C.

Enlace: <https://acortar.link/0G4hiG>.

Tema: Recibir realimentación y recomendaciones por parte de los miembros de la sala.

Hora: 2:00 p. m. a 4:00 p. m.

Fecha: 31/05/2024

Dependencia responsable: Secretaría Técnica del CASEN.

Participantes

Miembros de la Sala Especializada para la Modernización Tecnológica de la Producción Estadística del CASEN

Mario Linares Vásquez.

Nicolás Cardozo Álvarez.

León Darío Parra.

Departamento Administrativo Nacional de Estadística (DANE)

Julieth Alejandra Solano Villa, directora técnica de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN) y secretaria técnica del CASEN.

Luis Martin Barrera Pino, jefe Oficina de Sistemas (OSIS).

Natalia Ximena Arteaga Gutiérrez, asesora subdirección.

Andrea Catherine Neira Bustamante, designada OSIS.

Alejandro Sandoval Pineda, designado OSIS.

Daniel Mauricio Montenegro Reyes, designado OSIS.

Luis Fernando Barajas Duarte, designado OSIS.

Victoria Adriana González Ramírez, designada DSCN.

Juan José Rubio Mesa, designado DIMPE.

Secretaría Técnica del CASEN - DIRPEN

Julieth Alejandra Solano Villa, directora técnica DIRPEN y secretaria técnica CASEN.

Elizabeth Moreno Barbosa, asesora.

Yennifer Castillo Murcia, GIT Planificación y Articulación Estadística (PAE).

Sandra Yaneth Cortés Gamba, responsable general del CASEN. GIT PAE.

Derly Vivian Lizarazo García, responsable de la Sala Especializada para la Modernización Tecnológica de la Producción Estadística.

Mauricio Valencia Amaya, GIT PAE.

Agenda

Tiempo	Actividad	Responsable
2:00 p. m. a 2:05 p. m.	Instalación, verificación del quorum y registro fotográfico.	Derly Lizarazo, responsable de la sala.
2:05 p. m. a 2:10 p. m.	Apertura del evento.	Julieth Solano, directora técnica DIRPEN y secretaria técnica del CASEN.
2:10 p. m. a 2:15 p. m.	Síntesis de la reunión anterior y los compromisos.	Derly Lizarazo, responsable de la sala.
2:15 p. m. a 3:15 p. m.	Presentación de la línea de investigación 1. <i>Ciencia de datos, 1.1. Ciclo de vida de modelos de proyectos de ciencias de datos, No. 1.1. Revisión de prácticas de Ingeniera de Software para Machine Learning</i> y realimentación por parte de los miembros de la sala.	Moderador: Alejandro Sandoval, designado OSIS.
3:15 p. m. a 3:35 p. m.	Avances en la organización del seminario.	Derly Lizarazo, responsable de la sala.
3:35 p. m. a 3:40 p. m.	Compromisos de la reunión.	Derly Lizarazo, responsable de la sala.
3:40 p. m. a 3:55 p. m.	Conclusiones y cierre.	Julieth Solano, directora técnica DIRPEN y secretaria técnica del CASEN.

Objetivo

Recibir realimentación por parte de los miembros de la sala sobre la línea de *investigación 1. Ciencia de datos, 1.1. Ciclo de vida de modelos de proyectos de ciencias de datos, No. 1.1. Revisión de prácticas de Ingeniería de Software para Machine Learning* y conocer los avances en la organización del seminario.

Desarrollo

1. Instalación, verificación del quórum y apertura de la reunión

Se realizó la verificación del quórum, contando con la participación de los tres miembros de la sala, los designados de las Oficina de Sistemas, de la Dirección de Síntesis y Cuentas Nacionales (DSCN) y de la Dirección de Metodología y Producción Estadística (DIMPE), así como de la secretaria técnica del CASEN y su equipo de trabajo. Posterior a ello, se realizó el registro fotográfico y se le dio la palabra a Elizabeth Moreno, asesora de la dirección técnica de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN), quien realizó la apertura a la reunión.

2. Síntesis de la reunión anterior y compromisos

Derly Lizarazo, responsable de la sala, realizó una breve síntesis de la reunión anterior y señaló que el objetivo fue recibir realimentación por parte de los miembros de la sala sobre la línea de investigación No. 1.2. *Proyectos de ciencia de datos*, referente con el *índice de noticias* y presentación del artículo *Colombian agricultural sector's early estimator of Gross Domestic Production using Nowcasting and Big Data methods* y concertar la propuesta del evento de transferencia de conocimiento y de los demás productos de la sala.

También hizo referencia a los compromisos adquiridos: remitir material de trabajo y asunto de consulta para la línea de investigación *1. Ciencia de datos, 1.1. Ciclo de vida de modelos de proyectos de ciencias de datos, No. 1.1.1. Revisión de prácticas de Ingeniería de Software para Machine Learning*, elaborar y compartir acta de la reunión anterior y la reunión para acordar la agenda del seminario (17 de mayo) y avanzar en su preparación.

3. Presentación de la línea de investigación *1. Ciencia de datos, 1.1. Ciclo de vida de modelos de proyectos de ciencias de datos, No. 1.1. Revisión de prácticas de Ingeniería de Software para Machine Learning* y realimentación por parte de los miembros de la sala

La presentación la realizó Alejandro Sandoval, de la Oficina de Sistemas, y en ella relacionó 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 y se dio espacio para discutir sobre la metodología de ingeniería de software para Machine Learning que mejor se acople a las necesidades de la entidad respecto al desarrollo de código en proyectos de ciencia de datos. El asunto de consulta fue solicitar recomendaciones a los miembros de la sala sobre las prácticas de ingeniería de software para Machine Learning.

Para lo anterior, se propusieron a los miembros de la sala las siguientes preguntas:

Pregunta 1. En relación con la descripción de cada etapa de MLOps, ¿consideran que es necesario ajustar algún aspecto para mejorar la claridad y la comprensión de cada una de ellas?

Pregunta 2. En cuanto a los porcentajes de participación por rol en cada etapa, ¿consideran que se debe incluir algún rol adicional o ajustar los porcentajes existentes para reflejar mejor las contribuciones de cada uno?

Recomendaciones:

- Frente a la segunda pregunta de los roles adicionales, se recomendó incluir 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 posteriormente.
- En la validación falta participación de los *stakeholders*, ya que ellos son los que aceptan el trabajo que se está haciendo y es importante la validación temprana que pueden aportar.
- También se sugirió la inclusión de un ingeniero de desarrollo o de software ya que hará parte del equipo que va a desplegar y va a embeber el modelo dentro de un sistema más grande o el que le va a poner una interfaz más adecuada.
- La inclusión de los roles dependerá también de la complejidad. Sin embargo, es importante definir cuál sería 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 contarán con una clase para todos los modelos o los proyectos en la API.
- Para que el modelo sea reutilizable hay que publicarlo como una librería, donde hay una función con una definición clara de las entradas y las salidas.
- Se debe incluir a una persona que esté 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.

- Es importante documentar los modelos o dentro del código o 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.
- Debería existir un rol que verifique la calidad del dato y la seguridad de la información.

Pregunta 3. En relación con la estructura del árbol de codificación: ¿Consideran que su robustez es adecuada? ¿O resulta ser demasiado compleja o excesivamente simple? ¿Agregarían o eliminarían algún elemento para mejorarla?

Recomendaciones:

- La guía propuesta no debe ser una camisa de fuerza, sino un ente estructurador. El árbol propuesto está bien y dependerá de la complejidad del proyecto, donde se debe tener en cuenta el tema de la visualización y tener en cuenta un *notebook* adicional que indique el orden en que se debe proceder con lo respectivo.
- Se sugiere agregar un marco de principios y valores a la guía, y una metodología como *Extreme Programming*, que más o menos da lineamientos de cómo organizar el código.
- La guía propuesta debe tener la suficiente versatilidad para ajustarse dependiendo de las necesidades del equipo interno y la visión del usuario.
- Al construir el modelo es importante dejar documentado todo su proceso de experimentación, antes de llegar a los parámetros adecuados (llevar una bitácora de las pruebas realizadas).

Luego de recibir la realimentación sobre las preguntas, se revisó el gráfico elaborado para centralizar las herramientas que más se utilizan en el ámbito académico y empresarial y se discriminó en etapas. También se sugirió que, a partir del gráfico presentado, se puede montar un *massive open online courses* (MOOC) que en una primera versión tenga vídeo, animación o un contenido estático que presente la consolidación de herramientas que se pueden usar.

4. Avances en la organización del seminario “Arquitectura de datos: fundamentos y buenas prácticas”

Elizabeth Moreno, mencionó que el seminario puede ser un espacio valioso para la articulación con el Comité de Administración de Datos (CAD), como otra instancia de coordinación del SEN y se propuso la participación de MinTIC como consolidar de las conclusiones del evento, los miembros de la sala estuvieron de acuerdo.

Luego se mencionaron los documentos en los que se ha avanzado en la organización del seminario: agenda, nota conceptual, propuesta de correo de invitación, propuesta de invitaciones y *Forms* de inscripción. Asimismo, se presentó la pieza de invitación al seminario.

6. Compromisos de la reunión

Dado que en junio se realizará el seminario de seminario de *"Arquitectura de datos: fundamentos y buenas prácticas"*, y una reunión para presentación de los avances del PEN, se acordó retomar las actividades del plan de trabajo en julio (12 y 26).

Compromisos

Tarea	Compartir plantilla para el diligenciamiento de los datos de los invitados de la academia.
Responsable	Sandra Cortés Gamba, responsable general del CASEN.
Fecha entrega	04/06/2024.
Tarea	Compartir nota conceptual y pieza de invitación al seminario de <i>"Arquitectura de datos: fundamentos y buenas prácticas"</i> , para revisión de los miembros y participantes de la sala.
Responsable	Sandra Cortés Gamba, responsable general del CASEN.
Fecha entrega	07/06/2024.
Tarea	Remitir acta de la presente reunión para la revisión y la aprobación de los miembros y los participantes de la sala.
Responsable	Sandra Cortés Gamba, responsable general del CASEN.
Fecha entrega	11/06/2024.
Tarea	Remitir los correos de invitación para el seminario de <i>"Arquitectura de datos: fundamentos y buenas prácticas"</i> .
Responsable	Responsable de sala.
Fecha entrega	12/06/2024.
Tarea	Desarrollo del seminario de <i>"Arquitectura de datos: fundamentos y buenas prácticas"</i> .
Responsable	Secretarías Técnicas del CASEN y el CAD.
Fecha entrega	20/06/2024.

Próxima reunión

Responsable de convocar: Secretaría Técnica del CASEN.

Fecha: 12 de julio de 2024 de 2:00 p. m. a 4:00 p. m. (virtual).