Consejo Asesor Técnico del Sistema Estadístico Nacional (CASEN) 2023-2025

Sala Especializada para la Modernización Tecnológica de la Producción Estadística
ACTA No. 2

Ciudad: Bogotá D.C.
Enlace: https://acortar.link/mK5xmJ
Tema: Aprobación del plan de trabajo 2024-2025 e inicio de la línea de investigación No. 1.

Hora: 2:00 p. m. a 4:00 p. m.
Fecha: 22/03/2024
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 (E) de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN) y secretaria técnica CASEN.

Elizabeth Moreno Barbosa, asesora DIRPEN

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

Natalia Ximena Arteaga Gutiérrez, asesora subdirección.

Andrea Roncancio, coordinadora Grupo Interno de Trabajo (GIT) Prospectiva y Analítica de Datos (PAD), DIRPEN.

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

Yennifer Castillo Murcia, GIT – PAE Nacional

Sandra Yaneth Cortés Gamba, facilitadora general del CASEN.

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

Agenda

Tiempo	Actividad	Responsable
2:00 p. m. a 2:05 p. m.	Instalación, verificación del quorum y registro fotográfico.	Derly Lizarazo, facilitadora de la sala.
2:05 p. m. a 2:10 p. m.	Apertura de la reunión.	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 compromisos.	Derly Lizarazo, facilitadora de la sala.
2:15 p. m. a 2:25 p. m.	Presentación de los ajustes realizados al plan de trabajo de la sala.	
2:25 p. m. a 3:30 p. m.	Presentación de insumos de la línea de investigación No. 1. *Ciencia de datos,* 1.1. *Ciclo de vida de modelos de proyectos de ciencia de datos, actividad: definición del marco teórico asociado.*	Alejandro Sandoval, designado OSIS.
3:30 p. m a 3:40 p. m.	Comentarios adicionales.	Modera: Derly Lizarazo, facilitadora de la sala.
3:40 p. m a 4:00 p. m.	Compromisos, conclusiones y cierre.	Julieth Solano, directora técnica DIRPEN y secretaria técnica del CASEN.

Objetivo: aprobación del plan de trabajo 2024-2025 e inicio de la línea de investigación No. 1. *Ciencia de datos*, 1.1. *Ciclo de vida de modelos de proyectos de ciencia de datos, actividad: definición del marco teórico asociado.*

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 la 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 secretaría técnica del CASEN y su equipo de trabajo. Posterior a ello, se realizó el registro fotográfico y se le dio la palabra a Julieth Solano, directora técnica de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN) y secretaría técnica del CASEN, quien realizó la apertura a la reunión.

2. Síntesis de la reunión anterior y compromisos

Derly Lizarazo, facilitadora de la sala, realizó una breve síntesis de la reunión anterior, señalando que los temas tratados fueron referentes con el contexto y la conformación del CASEN, la importancia del rol de los miembros del CASEN y de las áreas técnicas del DANE, así como la metodología de trabajo, los resultados, los productos esperados y los planes de trabajo, tanto de la Sala especializada para la Modernización Tecnológica como para la Sala General.

Posterior a ello, recordó los compromisos establecidos concernientes con el envío del plan de trabajo a los miembros de la sala, según sus consideraciones y del suministro del material de trabajo para la presente reunión, y cumpliendo con las fechas programadas.

3. Presentación de los ajustes realizados al plan de trabajo de la sala

La facilitadora de la sala mencionó los ajustes realizados al plan de trabajo en las diferentes líneas de investigación y actividades, propuestas por los miembros del CASEN y que fueron aprobados. Frente a esto, el profesor Mario Linares propuso evaluar para los eventos de transferencia del conocimiento, involucrar a otros profesores y expertos externos al CASEN, con el fin de robustecer las temáticas. Al respecto, los otros dos profesores asintieron y mencionaron que ellos pueden contribuir con estos contactos. Al respecto, Julieth Solano, directora técnica de DIRPEN y secretaria técnica del CASEN, consideró el tema muy interesante y resaltó la importancia de dimensionarlo para revisar los aspectos logísticos. Posterior a ello, se le dio la palabra a Alejandro Sandoval, designado de la OSIS:

Alejandro Sandoval, inició con el asunto de consulta para los miembros de la sala, referente con establecer el marco conceptual (listado de conceptos con sus definiciones). Para lo anterior, presentó diferentes definiciones, entre ellas: ciencia de datos; aprendizaje de máquinas; inteligencia artificial; aprendizaje profundo; visión de computadora; procesamiento de lenguaje natural; reconocimiento

óptico de caracteres; reconocimiento automático del habla; automatización; canalización; *Development Operations* (DevOps), y *Machine Learning Operations* (MlOps).

Frente a lo expuesto, el profesor Mario Linares propuso la inclusión de otros conceptos como: ciclo de vida; IA generativa; IA general; modelos de lenguaje de gran tamaño (LLMs), IA responsable, modelo, aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por refuerzo y *Data cleaning*. Frente a lo anterior, Alejandro Sandoval estuvo de acuerdo.

Posterior a ello, continuó realizando una comparación de metodologías ágiles para ciencia de datos como: *Knowledge Discovery in Databases* (KDD), *Cross-Industry Standard Process for Data Mining* (CRISP-DM), Sample, Explore, Modify, Model and Assess (SEMMA) y *Team Data Science Process* (TDSP).

Al respecto, el profesor Mario Linares comentó que no consideraba las metodologías KDD y CRISP ágiles por ser lineales y burocráticas, contrario a la TDSP que cuenta con conceptos de agilismo incluidos. Por lo anterior, manifestó la importancia de hacer claridad en los conceptos a que se refiere con ágiles y compartió dos enlaces para conocer las metodologías *AN Analitycs*, *Guerrilla Analytics* y propuso tener en cuenta autores como Scott Ambler y Martin Fowler, que si son ágiles.

El profesor León Parra propuso tener en cuenta la metodología *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. La recomendación se enfocó en la importancia de combinar lo técnico y la gestión de proyectos, para dar agilidad en su implementación.

A continuación, Alejandro Sandoval presentó una diapositiva donde de manera gráfica se representaba la adaptación del TDSP y lo que se está manejando actualmente en el DANE para la ejecución de proyectos, no solo de ciencia de dato sino también de automatizaciones.

Frente a lo presentado, el profesor Nicolas Cardozo sugirió realizar un proceso de versionamiento para reconocer los avances y las diferencias presentadas. Igualmente, el profesor Mario Linares, asintió y complementó con la importancia de versionar la documentación y compartió un artículo sobre cómo aplicar la ingeniería de software para *machine learning*, con el fin de evaluar si se puede alinear con el TDSP.

Alejandro Sandoval agradeció los aportes y continuó exponiendo las etapas del TDSP, frente a lo que el profesor León Parra recomendó 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.

Al respecto, el profesor Mario Linares recomendó que en el documento de la arquitectura de datos se debe tener en cuenta si los datos van a un proceso de fusión, dado que de ser así se requiere incluir la programación, la implementación y el diseño de las fuentes. Asimismo, comentó que para evaluar

modelos se deben enfocar en cuatro atributos: ¿qué tan bueno es el modelo para la tarea?; el perfomance visto desde el rendimiento a consumir recursos; la seguridad, y el lidiar con los datos antagonistas.

Frente a lo anterior, Alejandro Sandoval expuso algunas de las herramientas que contribuyen con la limpieza de los datos. Al respecto, el profesor León Parra sugirió el uso de *bandax* que tiene una gran cantidad de librerías que combinándolas con *phyton* puede hacer más amigable la usabilidad de los modelos y la ayuda con la calidad de los datos. Al respecto, el profesor Mario Linares sugirió contactar a Tito Neira y a ingenieros de Augusta para conocer su experiencia.

Posterior a ello, Alejandro Sandoval presentó los roles establecidos en el TDSP (gerente de grupo, líder de equipo, líder de proyecto, ingeniero de datos, analista de datos, científico de datos e ingeniero de machine learning) y comentó que en el DANE existe la falencia porque el grupo es muy pequeño. Ante lo presentado, el profesor Mario Linares comentó que el ingeniero de *machine learning* debe implementar, desplegar y también validar.

Para finalizar la presentación, Alejandro Sandoval, expuso brevemente los proyectos desarrollados en el DANE, como automatización, big data, ciencia de datos y flujos de datos, y nuevamente agradeció a los miembros de la sala por sus valiosos aportes.

Dado que a lo largo de la presentación se realizó la realimentación y la discusión no se realizaron comentarios adicionales por parte de los miembros de la sala.

4. Compromisos, conclusiones y cierre

Derly Lizarazo,facilitadora de la sala, presentó los compromisos para la próxima reunión y dio la palabra a Julieth Solano, directora técnica de DIRPEN y secretaría técnica del CASEN, para las conclusiones y el cierre de la reunión, quien agradeció a Alejandro Sandoval y a los miembros de la sala por la interesante discusión y resaltó lo siguiente en cuanto a las conclusiones:

- Hacer ajustes a la inclusión de los conceptos planteados por los expertos como: ciclo de vida e inteligencia artificial generativa, inteligencia artificial, responsable analítica, aprendizaje supervisado, no supervisado por refuerzo, modelo y frente a estos elementos.

- Frente a las metodologías ágiles se recomendó revisar la pertinencia del CDM y metodologías como KD, así como aquellas que permitan combinar lo técnico con la gestión de los proyectos.

- Se recomendó realizar el versionamiento tanto del código como de los datos, el modelo, la documentación e incluso de la implementación. Al respecto, el experto Mario Linares compartió en

el chat una serie de documentos relacionados con ingeniería de software y *machine learning* para una posterior revisión.

- Respecto a la limpieza de los datos, el experto Nicolás Cardozo recomendó incorporar tareas que reconozcan o consideren métricas sobre la calidad de esta limpieza y la efectividad. Por su parte, el experto Mario Linares considera que este tema también se debe enfocar a la adquisición y el entendimiento de los datos, es decir hacia la arquitectura de los datos y no solo la arquitectura de la solución. El experto León Darío mencionó la importancia de los elementos asociados con la estandarización de los procesos de limpieza de datos a partir de considerar indicadores de consistencia e integridad.

- Frente a la etapa de modelamiento, la recomendación del experto Mario Linares se refirió a considerar dos niveles de la arquitectura, un elemento conceptual y un elemento físico y el evaluar los modelos considerando que tan bueno es el *performance* o el rendimiento.

- Se presentó la propuesta de conversar con el experto Tito Neira, para aprender o tener algunos elementos del proceso que se ha seguido en el marco de la iniciativa de Augusta.

- Se recomendó 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 permita hacer una gestión más efectiva del conocimiento.

Compromisos

Tarea	Envío de acta de la presente reunión para revisión y aprobación de los miembros y participantes de la sala.
Responsable	Facilitadora de la sala: Derly Lizarazo.
Fecha entrega	02/04/2024.
Tarea	Remitir el material de trabajo y el asunto de consulta para la próxima reunión, referente con la línea de investigación No. 1: Ciclo de vida del modelo de proyecto de ciencia de datos: 1.1. Documento: "*Resumen de los aspectos teóricos y conceptuales relacionados con el ciclo de vida de modelos en proyectos de ciencia de datos para la correspondiente revisión de los miembros expertos del CASEN*". 1.2. Proyectos de ciencia de datos.
Responsable	Facilitadora de la sala: Derly Lizarazo.
Fecha entrega	12/04/2024.
Tarea	Revisión del material de trabajo y recomendaciones al respecto.
Responsable	Miembros de la sala.
Fecha entrega	26/04/2024.

Tarea	Avanzar en la propuesta para el evento de transferencia del conocimiento.
Responsable	Miembros y participantes de la sala, con apoyo de Derly Lizarazo, facilitadora de la sala.
Fecha entrega	26/04/2024.

Próxima reunión

Responsable de convocar: Secretaría Técnica del CASEN.

Fecha: 26 de abril de 2024 de 2:00 p. m. a 4:00 p. m. (presencial en el DANE).