

ACTA DE REUNIÓN N°2 AYUDA DE MEMORIA

Ciudad: Bogotá

Lugar: Sesión virtual, plataforma Teams

Tema: Presentación proyectos DANE y aportes expertos

Hora: 8:30 a.m. a 10:40 a.m.

Fecha: 9/06/2021

Dependencia responsable: Secretaría Técnica CASEN (DIRPEN)

Participantes

Miembros de la Sala Especializada de Modernización Tecnológica

León Darío Parra Bernal
Jorge Andrés Gallego Durán
Valérie Gauthier Umaña

Asesor de la Dirección General
Juan Sebastián Ordoñez

Dirección general
Liliana Morales Hurtado

Prospectiva y Análisis de Datos
Alexander González

Calidad Estadística
María del Pilar Gómez

Prospectiva y Análisis de Datos
Raúl Andrés Gómez

Prospectiva y Análisis de Datos
Grace Andrea Torres

DIRPEN-Contratista
Mateo Cardona

Profesional GIT Planificación y Articulación Estadística (DIRPEN)
María Laudice Barreto

Profesional GIT Planificación y Articulación Estadística (DIRPEN)
María del Pilar Pulido

Departamento Administrativo Nacional de Estadística (DANE)

Coordinadora (E) Planificación y Articulación Estadística
Elizabeth Moreno Barbosa

Oficina de Sistemas
Marly De Moya Amaris

Oficina de Sistemas
Karin Stefany Muñoz

Orden del día

1. Solicitud autorización para grabar la reunión
 2. Verificación del quórum
 - I. León Darío Parra Bernal
 - II. Valérie Gauthier Umaña
 - III. Jorge Andrés Gallego Duran
 - IV. Elizabeth Barbosa
 - V. Marly Esther De Moya Amaris
 - VI. Juan Sebastián Ordoñez
 3. Síntesis reunión 12 mayo (Pilar Pulido - 2 minutos)
 4. Presentación Desarrollo y Distribución de Aplicaciones Web para la Recolección de Datos (Raúl Andrés Gómez – 10 minutos)
 5. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)
 6. Detección de Anomalías Temáticas y Operativas en el Censo Económico (Raúl Andrés Gómez – 10 minutos)
 7. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)
 8. Medición proxy del indicador del ODS 10.3.1/16.b.1 utilizando fuentes no convencionales (redes sociales) (Grace Andrea Torres – 10 minutos)
 9. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)
 10. Conclusiones y cierre (Elizabeth Barbosa - 5 minutos)
-

Desarrollo

Objetivo

Presentar los proyectos desarrollados por el DANE, a fin de recibir realimentación por parte de los expertos y fortalecer estos procesos.

1. Solicitud autorización para grabar la reunión

Se inició la grabación de la sesión con previa autorización de los participantes.

2. Verificación del quórum

Se verificó el quórum tanto de los expertos de la Sala, como del Departamento Administrativo Nacional de Estadística (DANE).

3. Síntesis reunión 12 mayo (Pilar Pulido - 2 minutos)

Pilar Pulido inició la reunión e indicó como propuesta de valor para el presente año proponer nuevos temas en analítica y seguridad, promover productos de divulgación y trabajar en un documento que recoja los aportes del trabajo realizado en las reuniones de la sala, así como la generación de espacios tipo webinar para temas de seguridad en la información y del ecosistema de datos en cuanto a producción estadística.

Pilar Pulido mencionó que, de acuerdo a las recomendaciones de expertos a nivel estratégico y táctico en la línea de interoperabilidad, se identificaron los avances realizados por el DANE, las tareas por realizar y los aportes esperados por parte de los expertos de la Sala.

Asimismo, se realizó la presentación de la Política de Gobierno de RRAA y Fuentes Alternativas, la Política de Seguridad de la Información del DANE, el modelo Gestión de Datos y requerimientos de seguridad, así como el estado de avance de los proyectos implementados a través de la PDI.

En cuanto a los pilotos en el lago de datos local, se realizó la presentación de los Proyectos de educación D4N, anonimización –CEED y DSCN sector carbón y cruces difusos en PILA y DEST. Como potenciales pilotos, la automatización de la actividad económica CIU Rev. 4 AC, la detección de anomalías - Censo Económico y la automatización de procesos de calidad – Temática e Imputación de datos con cruce difuso – GEIH. En cuanto al lago de datos en Nube se indicaron como potenciales pilotos la detección de anomalías - Censo Económico, los indicadores sector carbón y pobreza monetaria multidimensional y ODS. Luego de dicha presentación y con los aportes de los expertos se acordó enviar insumos a expertos de proyectos DANE previo a las reuniones.

4. Presentación Desarrollo y Distribución de Aplicaciones Web para la Recolección de Datos (Raúl Andrés Gómez – 10 minutos)

Raúl Gómez comenzó la presentación indicando que el proyecto “Desarrollo y distribución de aplicaciones web para la recolección de datos” tiene como objetivo desarrollar y distribuir, por medio de una red social, una aplicación web, como método alternativo de recolección de información estadística

oficial; como alcance 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).

Como justificación señaló que la recolección de información por encuestas es cada vez un reto mayor para los INES, adicional se ha identificado que las personas gastan cada vez más tiempo en el uso de dispositivos electrónicos y redes sociales y gran parte del contenido de internet lo generan los usuarios UGD. Por otro lado, los algoritmos de las plataformas de redes sociales permiten la distribución rápida y dirigida 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 de encuesta. De acuerdo a lo anterior señaló que el colombiano promedio usa dispositivos y redes 3 horas y el al día y el 63% usa Facebook, de este porcentaje, el 97,8 que usan Facebook lo hacen con dispositivos móviles.

Señaló como productos del presente proyecto: un documento metodológico, la aplicación web y conclusiones sobre el medio y método de recolección, así como datos recolectados. Frente a los resultados se buscará reducción de costos y tiempo de recolección de información, acceso a nueva información, evitar carga a los encuestados y mejorar la calidad de la información (Información orgánica).

En cuanto al cronograma se estableció el desarrollo de las siguientes actividades para un periodo comprendido entre mayo y diciembre del presente año: especificación de necesidades de información; diseño de variables y modelo de datos; diseño o descripción de la estrategia de recolección; diseño estadístico; diseño de propuesta de valor y lógica de la aplicación web; diseño de la experiencia del usuario, ética y cuestiones legales; diseño, desarrollo y lanzamiento de la aplicación web; distribución aplicación web y recopilación de datos; procesamiento, análisis y difusión de los resultados finales y evaluación de los costos y beneficios del proceso y conclusiones de viabilidad del método.

De acuerdo a la exposición realizada, se indicó que, se busca obtener aportes de manera puntual por parte del experto León Darío Parra frente a capacitaciones y gestión identificadas en temas de desarrollo No Code de aplicaciones web, diseño ético de aplicaciones, experiencia de usuario o comportamiento de usuario o economía de comportamiento y servicios de Facebook.

5. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)

León Darío Parra inició su presentación indicando que la contribución al proyecto anteriormente descrito tendría como objetivo contribuir y apoyar con el asesoramiento técnico al diseño e implementación del PMV de la aplicación para recolección de datos en redes sociales o fuentes alternativas no estructuradas.

El alcance que se puede establecer es contar con el apoyo de dos profesores más de la Universidad EAN quienes se unirían al equipo técnico asesor (50 Horas) para el diseño de la aplicación orientada al usuario. (Posibilidad de invitar a un Experto en Habeas Data y Uso de nuevas Tecnologías – Ético) y la metodología propuesta se enfocaría en el diseño del proyecto bajo la metodología SCRUM.

Contempla tener en cuenta un mapa de experiencias donde se vincule “evolution, applications y emerging research” donde primero se realizan desarrollos de estructura con aplicaciones web con diferentes interfaces y, se involucra web analytics para el análisis con redes neuronales. Indicó también la importancia del uso de metodologías ágiles centradas en el usuario donde se inicia con la percepción del usuario y se finaliza con las versiones del proyecto, en este sentido, el aporte sería facilitar el proceso con asesoría o apoyo, porque se trabaja en función del usuario final, con 3 o 4 interacciones con el PMV, en función del listado de requerimientos.

Así mismo, señaló que en función de la muestra de usuarios las aplicaciones dan los resultados con respecto al alcance que se espera dar, con Twitter por ejemplo, dan dos opciones, una opción de muestra gratuita, pero para medir hipótesis o correlaciones el nivel de significancia es muy acotado, para subsanar, se ofrecen paquetes con un cobro.

Finalmente realizó dos propuestas; la primera realizar un ejercicio artesanal usando como ejemplo el desarrollo de una App para el diagnóstico de la anamnesis, teniendo en cuenta sus limitaciones, y la segunda propuesta enfocada en el uso de herramientas estandarizadas como Cielo Go, que agrupa empresas del Big Data para este tipo de aplicaciones y algoritmos.

Continuando con el ejercicio de aportes por parte de los expertos de la sala, Jorge Gallego indicó que está un poco confundido frente a la necesidad del objetivo planteado frente a si se tiene que desarrollar una aplicación, pues ya hay herramientas como SurveyMonkey para realizar este tipo de encuestas on line a unos costos más bajos, considerando los recursos asociados al desarrollo de una aplicación nueva. Referente al muestreo indicó que se pueden presentar diferencias entre muestra y población, y es acá donde se pueden ver dos caminos, el primero donde no se busca muestra representativa sino población objetivo y el segundo, reconociendo que si hay interés por una muestra representativa y se realizan ajustes estadísticos.

Valérie Gauthier mencionó que tiene duda frente al valor que da la aplicación y pregunta si es suficiente, pues hay que considerar el feedback como un gana gana, esto debe ser lo que se busca. Adicional, mencionó qué cambiaría en el cronograma, empezaría con la especificación de necesidades de información, luego con el diseño de propuesta de valor, y luego con el diseño de la experiencia del usuario. También indicó sobre el por qué compartir en redes sociales, por qué se debe descargar una aplicación, y si dicha aplicación tendrá un cuestionario, pues puede ser un poco engorroso para los usuarios el proceso. Indicó que, si facebook sería el gran aliado, este constituiría un respaldo serio. Finalmente señaló que se puede presentar negación por parte del usuario a contestar encuestas del DANE y por esta razón se debe buscar una solución frente a los resultados obtenidos como por ejemplo un acompañamiento a los encuestados.

Luego de la intervención de los expertos, Elizabeth Barbosa comentó que se buscará la manera cómo conciliar las propuestas y comentarios realizados para realizar ajustes al proyecto.

6. Detección de Anomalías Temáticas y Operativas en el Censo Económico (Raúl Andrés Gómez – 10 minutos)

Raúl Gómez comenzó su intervención señalando que el objetivo del proyecto “Detección de Anomalías Temáticas y Operativas en el Censo Económico” consiste en Entrenar modelos de aprendizaje automático que permitan identificar anomalías en el proceso operativo de recolección de datos y en la información reportada por las fuentes del Censo Económico 2022. Teniendo en cuenta lo anterior, se planteó como alcance identificar anomalías operativas (ejecución de la entrevista) y temáticas (consistencia de los valores de las variables), detectar anomalías durante la fase de recolección de los datos del Censo Económico que permitan la corrección de ellos datos en campo teniendo en cuenta los sectores de Industria, Comercio, Servicios y los sectores de la encuesta de Micronegocios que son cubiertos por el Censo, así como generar indicadores y resultados agregados para el Sistema de Monitoreo y Control del Censo Económico, para esto durante 2021 se realizarán pruebas de campo y se finalizará el diseño para su implementación en 2022.

En cuanto a la justificación del proyecto, indicó que el Censo Económico es una operación estadística poco frecuente, grande, compleja y costosa, lo que hace que la gestión de la calidad de los datos sea más desafiante y requiere métodos robustos; los métodos de ML pueden ser más precisos y se pueden automatizar, contribuyendo a la eficacia y eficiencia en la revisión de la calidad de los datos. Tradicionalmente se emplean métodos univariados en el análisis de anomalías, los métodos basados en el ML incorporan las relaciones de múltiples variables, detectando anomalías que los métodos univariados no detectan. Las reglas de validación son dependientes del conocimiento temático y no consideran relaciones en los datos que no son conocidas previamente, algunas relaciones conocidas entre los datos son costosas computacionalmente y en programación, finalmente, indicó que la detección de anomalías contribuye a evitar problemas en los atributos de precisión, exactitud y coherencia de los datos y su consecuente utilidad para los usuarios.

Raúl Gómez indicó que a partir de los productos del proyecto se esperan obtener modelos de detección de anomalías, métricas de desempeño del modelo seleccionado, documento de desarrollo y scripts de desarrollo. En cuanto a los resultados se busca una reducción de tiempos y costos de seguimiento y control, así como control de la calidad de los resultados estadísticos. Referente a los impactos, se busca generar mayor confianza en los resultados (coherencia, exactitud), mejora de la solidez metodológica del DANE, eficiencia y modernización de los procesos y que sea escalable y reproducible a otras OOE.

Para desarrollar dicho proyecto se establece un cronograma que vincula las fases de diseño, construcción, prueba, ajustes al diseño y la construcción, así como su implementación, la cual se tiene proyectada para el mes de noviembre del 2022.

Como avances, Raúl Gómez presentó las fases de entendimiento del problema, entendimiento de datos y preparación de datos como terminadas; las fases de modelado y evaluación en proceso y la fase de implementación sin iniciar.

Finalizando su intervención mencionó los aportes que se pueden esperar de los expertos Valérie Gauthier Umaña y Jorge Andrés Gallego Durán frente a los retos identificados: datos recolectados con fines estadísticos, sesgo de variables de interés, representatividad, adaptación del dominio, múltiples dominios y dimensionalidad.

7. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)

Valérie Gauthier señaló que, de acuerdo con la conversación con Juan Fernando Pérez Bernal; por tratarse de un equipo interno, es mejor hacer procesos iterativos, probar, mirar resultados, volver a probar, etc. Hay que escoger el modelo, pero con análisis exploratorio de los datos. Para esto se debe empezar por datos más sencillos, luego ver si es viable y si es replicable, se debe tener en cuenta la explicación del modelo, justificar el por qué los datos que no se tienen en cuenta. Adicional, el modelo debe estar escogido con base en la necesidad.

Indicó que sí se cuenta con proyectos más experimentales y con menos recursos se puede trabajar con estudiantes de últimos semestres para trabajar en conjunto.

Continuó con el desarrollo del punto Jorge Gallego, quien señaló que las universidades en algunos programas tienen trabajo profesionalizante para graduarse y se puede revisar la forma de realizar este trabajo con el DANE para brindar apoyo. En cuanto al proyecto expuesto, señaló que no es muy claro a dónde lleva la detección de las anomalías, quién es el consumidor final y qué se hace frente con las anomalías. Adicional indicó que se debe empezar de lo simple y llegar a lo complejo y que se debe tener cuidado en que no siempre anomalía es lo mismo que error, y es importante determinar las acciones a desarrollar frente a la anomalía, y es en este punto donde la parte humana es fundamental. Señaló que se debe tener en cuenta modelos como Local Outlier Factor que se ha utilizado para las bombas de gasolina en el caso puntual de Bogotá.

8. Medición proxy del indicador del ODS 10.3.1/16.b.1 utilizando fuentes no convencionales (redes sociales) (Grace Andrea Torres – 10 minutos)

Grace Torres inició su presentación al indicar que el objetivo del proyecto “Medición proxy del indicador ODS 10.3.1/16.b.1 utilizando fuentes no convencionales (redes sociales)” es obtener mediciones complementarias de los indicadores del ODS 16 asociados a la percepción utilizando fuentes no convencionales (redes sociales), en particular para los indicadores 16.7.2 Proporción de la población que cree que la toma de decisiones es inclusiva y receptiva, por sexo, edad, discapacidad y grupo de población y 10.3.1/ 16.b.1 Proporción de la población que declara haberse sentido personalmente discriminada o acosada en los últimos 12 meses por un motivo de discriminación prohibido por el derecho internacional de los derechos humanos. Para esto señaló que se entiende por discriminación *“toda distinción, exclusión, restricción o preferencia u otro trato diferenciado basado directa o indirectamente en los motivos expresamente prohibidos y que tiene la intención o el efecto de anular o dificultar el reconocimiento, el*

disfrute o el ejercicio en condiciones de igualdad por todas las personas de todos los derechos humanos civiles, culturales, económicos, políticos y sociales", y que el estudio dará cuenta de la discriminación directa, la cual "se produce cuando una persona recibe un trato menos favorable que otra por una razón relacionada con uno de los motivos prohibidos y sin justificación razonable y objetiva (por ejemplo, no se entrevista a una persona con calificaciones iguales o superiores debido a su origen étnico)" (Naciones Unidas).

Señaló que se presentan las siguientes formas de discriminación definidas por el metadato:

- SEXO: como ser mujer u hombre.
- EDAD: como ser percibido como demasiado joven o demasiado viejo.
- DISCAPACIDAD O ESTADO DE SALUD: como tener dificultades para ver, oír, caminar o moverse, concentrarse o comunicarse, que tiene una enfermedad u otras afecciones de salud y no se le han proporcionado ajustes razonables.
- ORIGEN ÉTNICO, COLOR O LENGUAJE: como el color de la piel o la apariencia física, el origen étnico o la forma de vestirse, la cultura, las tradiciones, el idioma nativo, el estado indígena o el ser de ascendencia africana.
- ESTADO DE MIGRACIÓN: como nacionalidad u origen nacional, país de nacimiento, refugiados, solicitantes de asilo, estatus migratorio, inmigrantes indocumentados o apátridas.
- CONDICIÓN SOCIOECONÓMICA: como la riqueza o el nivel educativo, percibido como perteneciente a un grupo o clase social o económica más baja o diferente, propiedad de la tierra o la casa o no.
- UBICACIÓN GEOGRÁFICA O LUGAR DE RESIDENCIA: como vivir en zonas urbanas o áreas rurales, asentamientos formales o informales.
- RELIGIÓN: como tener o no una religión o creencias religiosas.
- ESTADO CIVIL Y CONDICIÓN FAMILIAR: como ser soltero, casado, divorciado, viudo, embarazada, con o sin hijos, huérfano o nacido de padres solteros.
- ORIENTACIÓN SEXUAL O IDENTIDAD DE GÉNERO: como ser atraído por una persona del mismo sexo, autoidentificarse de manera diferente al sexo asignado al nacer o ser sexualmente, corporal y / o de género diverso.
- OPINIÓN POLÍTICA: como expresar opiniones políticas, defender los derechos de los demás, ser miembro o no de un partido político o sindicato.
- OTROS MOTIVOS

Como alcance del proyecto, se estableció definir la línea de base para medir la percepción de la discriminación mediante el uso de nuevas fuentes (datos de medios sociales) de información, así como analizar la percepción de la discriminación a través de temas según el marco de la ONU, definiendo un documento metodológico sobre la recopilación, el procesamiento y el análisis de los datos, impulsado por técnicas avanzadas de aprendizaje automático para el procesamiento del lenguaje natural. También se busca difusión de resultados y experiencias del equipo (GitHub y página web de ejercicios experimentales) en el uso de estas técnicas para la socialización de buenas prácticas con socios internacionales.

La justificación planteada para este proyecto establece que la promesa de no dejar a nadie atrás y eliminar la discriminación está en el centro de la Agenda 2030 para el Desarrollo Sostenible. La eliminación de la discriminación también está consagrada en la Declaración Universal de los Derechos Humanos y en los principales tratados internacionales de derechos humanos. El propósito de este indicador es medir la prevalencia de discriminación basada en la experiencia personal reportada por los individuos. Se considera un indicador de resultado que ayuda a medir la eficacia de las leyes, políticas y prácticas no discriminatorias para los grupos de población interesados (Metadato del Indicador 16.b.1., 10.3.1. Oficina del OACNUDH). Los Estados tienen la obligación de recopilar y publicar datos desglosados por motivos de discriminación en el derecho internacional de los derechos humanos. Entre estas categorías figuran el sexo, la edad, la etnia, la condición migratoria o de desplazamiento, la discapacidad, la religión, el estado civil, el ingreso, la orientación sexual y la identidad de género (Enfoque de datos basados en derechos humanos de la Oficina del ACNUDH). El Comité insta al Estado parte a llevar a cabo investigaciones exhaustivas de los casos de discriminación racial y sancionar debidamente a los responsables (Observaciones finales sobre los informes periódicos 17º a 19º combinados de Colombia de la Convención Internacional sobre la Eliminación de Todas las Formas de Discriminación Racial (CERD)).

Para la definición de los pasos a seguir, indicó que la base de este ejercicio es el análisis semántico, pues este permite identificar el significado de un término a través de diferentes técnicas, métodos y modelos en los que se estudia el término y las relaciones semánticas entre él y otros términos asociados (familia de conceptos), el objetivo es hacer un análisis de tipo no supervisado, en el que definimos unos términos de búsqueda e identificamos un conjunto de entrenamiento para poder entrenar un modelo que "aprenda" a identificar el concepto en ciertos contextos y lo asocie a la discriminación, para ello, debemos generar un campo semántico sobre el concepto "discriminación" (esto es, una especie de ontología).

Como avances señaló que las actividades de revisión preliminar de modelos y algoritmos basados en experiencias internacionales y preprocesamiento de datos se encuentran en proceso; las de definición de la metodología y del diseño experimental y recolección de datos y evaluación de la calidad de datos se encuentran terminadas y las de modelado, análisis de rendimiento del modelo, documentación de resultados preliminares, evaluación del modelo entrenado y documentación final (scripts, notebook, dashboard) no han comenzado aún.

Finalmente comentó que, como aportes a recibir por parte de los expertos de la sala, se enfocarán en los retos de restricciones de acceso a los datos, falta de capacidad de almacenamiento y procesamiento, baja representatividad de datos para el entrenamiento y especificidad de la temática.

9. Presentación expertos aportes solicitados para el proyecto (Expertos de la sala – 10 minutos cada uno)

Valérie Gauthier señaló que referente al proyecto expuesto tiene como pregunta cómo realizaron la ingesta de datos y que es interesante revisar el ejercicio con Facebook para tener información adicional. Adicional indicó que quedan atentos para acompañamiento y si tienen proyectos pendientes se pueden apoyar para trabajar con universidades.

Jorge Gallego comentó sobre una la duda acerca de si sólo se quiere realizar algo descriptivo, o si se van a tomar acciones de política pública; recomienda la lectura de Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. También señaló que no hay claridad si el método a utilizar es de aprendizaje no supervisado, y por lo tanto valdría mirar métodos semi supervisados.

León Darío Parra manifestó que sería interesante conocer los apoyos puntuales en los que como expertos pueden aportar, para no entrar en dinámicas de reproceso; para esto, se hace necesario el análisis exploratorio descriptivo para empezar con los algoritmos, y sería oportuno conocer las hipótesis y la pregunta con el respectivo algoritmo y a lo que se esperaría llegar con el resultado final, para determinar el mejor método. Finalmente señaló que el uso de metodologías ágiles permite tener iteraciones y proceso para resultados finales.

10. Conclusiones y cierre (Alexander González - 5 minutos)

Dado que la reunión se extendió en 10 minutos, no se alcanzó a realizar este punto

Compromisos

- 1. Tarea:** enviar el acta de la sesión
Responsable: DANE
Fecha entrega: a más tardar 17 de junio
- 2. Tarea:** envío Doodle para agendar tercera reunión
Responsable: DANE
Fecha entrega: 15 junio
- 3. Tarea:** envío documentación proyectos a trabajar en la sala
Responsable: DANE
Fecha entrega: antes de la realización de la tercera reunión

Próxima reunión:

Responsable de convocar: Secretaría Técnica CASEN
Fecha: Por confirmar