

Planteamiento del problema de Analytics – Adopción digital en Colombia

Para generar un entendimiento profundo de la adopción digital (AD) en Colombia y brindar insights accionables para las entidades gubernamentales, este proyecto tiene como objetivo realizar un perfilamiento detallado de la población en la que la adopción digital es baja. La meta es crear una herramienta de visualización que facilite este análisis, fundamentada en dos enfoques analíticos que exploren variables clave como la tasa de penetración y el acceso digital. Se implementarán dos soluciones: primero, un análisis de componentes principales (PCA) para identificar covariantes relacionadas con la adopción digital; y segundo, un modelo de clustering para segmentar la población en grupos con características específicas, orientando así el diseño de estrategias de política pública. El resultado final será un dashboard que resuma los hallazgos más relevantes para la toma de decisiones.

La información provendrá de encuestas del CNC y Saber 11, las cuales permitirán un perfilamiento detallado de los hogares y un análisis del nivel de apropiación digital en diversos segmentos demográficos. Además, se utilizarán datos no estructurados de reportes del DANE e ICFES para profundizar en el contexto socioeconómico de la conectividad en Colombia. Se emplearán técnicas de aprendizaje no supervisado, como PCA y clustering, con métricas de calidad de modelo (varianza explicada y cohesión de clusters), interpretabilidad y relevancia en políticas públicas, una varianza explicada alta permitirá el adecuado dimensionamiento de las características de los individuos, los que concederá que las agrupaciones sean heterogéneas entre si. También se definirán métricas de referencia basadas en estudios previos y datos oficiales. Los resultados ofrecerán al CNC una visión integral de la dinámica de la adopción digital en Colombia, proporcionando bases sólidas para la formulación de políticas que promuevan la inclusión digital.

Durante dos meses, un equipo compuesto por un Líder de Proyecto, un Científico de Datos, un Analista de Datos Senior y un Analista de Visualización trabajará en el desarrollo de esta solución analítica. El Líder de Proyecto coordinará al equipo y alineará las actividades con los objetivos estratégicos, mientras que el Científico de Datos desarrollará los modelos analíticos. El Analista de Datos Senior gestionará la transformación de los datos provistos por los stakeholders, y el Analista de Visualización se encargará de presentar los resultados de manera clara y accionable. Con un presupuesto de \$42,393,000 COP, el proyecto contará con recursos tecnológicos de procesamiento en la nube y herramientas como Jupyter Notebook y Power BI para el análisis y visualización de resultados.

Anexo.

Parte I. Del Qué del problema de negocio al Cómo del problema de Analytics

Fase a) Los encargados de la solución

Responde los siguientes planteamientos en una primera iteración con tu equipo de trabajo:

1. ¿Qué resultados se esperan obtener tras la implementación del proyecto?
Tras la implementación del proyecto se espera mejorar la identificación y perfilamiento de las características de los individuos con baja adopción digital. Esto a su vez facilitará la capacidad de generar intervenciones gubernamentales efectivas en poblaciones vulnerables, alineadas a la reducción de brechas digitales y la mejora de la apropiación digital de la región Colombia. Además, a partir de la implementación del proyecto se espera una optimización de recursos para intervenciones futuras basadas en datos analíticos y la creación de un marco analítico para monitorear y evaluar futuras iniciativas de inclusión digital.
2. ¿Quiénes participarán en la solución analítica?
En el desarrollo de la solución analítica participarán 4 recursos, incluyendo un líder de proyecto, un científico de datos, un analista de datos senior y un analista en visualización de datos. Luego, quién utilizará la solución analítica corresponde al Centro Nacional de Consultoría (CNC), que brindará un perfilamiento adecuado de las comunidades vulnerables afectadas por la brecha digital a las entidades gubernamentales correspondientes para que puedan ejecutar las acciones pertinentes de forma eficiente en pro de mejorar el nivel de apropiación digital de Colombia.
3. ¿Qué cambios se generarían en la organización/unidad cliente como resultado del proyecto?
Como resultado del proyecto se espera que el Centro Nacional de Consultoría (CNC) pueda comprender las características de la población vulnerable de Colombia en aspectos de adopción digital, así como la dinámica de su comportamiento en términos de variables explicativas como penetración y acceso. Este entendimiento permitirá que el CNC pueda proponer iniciativas junto a las entidades gubernamentales para reducir la brecha digital en Colombia.

Fase b) Lo que el cliente quiere

Con base en la herramienta “Análisis de Kano” u otra herramienta que el equipo elija con el propósito de generar información acerca de lo que quieren o esperan los clientes a diferentes niveles, así como priorizar estos requerimientos, respondan en equipo lo siguiente. En los contextos donde se cuenta con los *stakeholders* (SH) estas respuestas se buscan con ellos. Una consideración básica e importante es que todos los requerimientos deben ser **medibles**.

1. ¿Cuáles son los requerimientos básicos u obligatorios que se atienden con el proyecto?
 - El proyecto debe ayudar a CNC a generar un entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia.
 - En la solución del proyecto se deben utilizar variables explicativas como penetración y acceso.
 - El proyecto debe generar al menos dos mecanismos analíticos diferentes apoyados en modelos tipo *White, Gray o Black-Box Model*.
 - Se debe generar una herramienta de visualización de los modelos de la solución.
2. ¿Cuáles son aquellos requerimientos que se consideran indiferentes en el producto del proyecto?
 - La herramienta de visualización de los modelos de la solución se debe crear con PowerBI.
 - La solución del proyecto se debe disponibilizar como un servicio en la nube.
 - El desarrollo de los mecanismos analíticos del proyecto se debe desarrollar en Python.
 - Los insumos de datos del proyecto se deben alojar en los servidores del stakeholder.
3. ¿Cuáles son aquellos requerimientos considerados inversos?
 - La herramienta de visualización debe tener varias ventanas con la mayor información posible.
 - Los mecanismos analíticos desarrollados deben dividir a la población objetivo en muchos subgrupos (al menos 5).
 - El acceso a la solución analítica debe ser gestionado por el stakeholder. (Puesta en producción y disponibilización).
 - La solución analítica solo se puede usar para los datos otorgados por el stakeholder y no por datos nuevos.
 - La solución debe utilizar proveedores que involucren costos extra.
4. ¿Cuáles son los requerimientos unidimensionales?
 - Porcentaje de la población objetivo-perfilada. Mientras más personas con baja adopción digital identificada, las acciones a tomar por las entidades gubernamentales serán más precisas.
 - Rapidez de carga y proceso de los mecanismos analíticos.
 - Rapidez de carga y proceso de la herramienta de visualización.

- Número de insights accionables a partir del perfilamiento. Mientras más insights se obtengan del perfilamiento, más iniciativas se pueden crear a favor de reducir la brecha digital.
- Nivel de automatización de la solución. Mientras más cómoda sea la experiencia para el stakeholder, mejor.
- Reducción del tiempo de implementación de la solución en los sistemas del stakeholder.

5. ¿Qué requerimientos son atractivos?

- Mecanismos analíticos de la solución involucran Inteligencia Artificial Generativa.
- Paleta de colores corporativa del stakeholder utilizada en la solución.
- Horas de capacitación para los usuarios de negocio que consumirán la solución.
- Alertas en tiempo real de insights accionables.
- Personalización de la solución después de un año de prueba.

Fase c) Los requerimientos finales

Responde en equipo los siguientes planteamientos en una segunda iteración, con base en las respuestas encontradas en la fase b):

1. ¿Cuáles son los requerimientos? Enliste acorde a su clasificación.

- Básicos
 - o El proyecto debe ayudar a CNC a generar un entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia.
 - o En la solución del proyecto se deben utilizar variables explicativas como penetración y acceso.
 - o El proyecto debe generar al menos dos mecanismos analíticos diferentes apoyados en modelos tipo *White, Gray o Black-Box Model*.
 - o Se debe generar una herramienta de visualización de los modelos de la solución.
- Indiferentes
 - o La herramienta de visualización de los modelos de la solución se debe crear con PowerBI.
 - o La solución del proyecto se debe disponibilizar como un servicio en la nube.
 - o El desarrollo de los mecanismos analíticos del proyecto se debe desarrollar en Python.
 - o Los insumos de datos del proyecto se deben alojar en los servidores del stakeholder.
- Inversos
 - o La herramienta de visualización debe tener varias ventanas con la mayor información posible.
 - o Los mecanismos analíticos desarrollados deben dividir a la población objetivo en muchos subgrupos (al menos 5).
 - o El acceso a la solución analítica debe ser gestionado por el stakeholder. (Puesta en producción y disponibilización).
 - o La solución analítica solo se puede usar para los datos otorgados por el stakeholder y no por datos nuevos.
 - o La solución debe utilizar proveedores que involucren costos extra.
- Unidimensionales
 - o Porcentaje de la población objetivo perfilada. Mientras más personas con baja adopción digital identificada, las acciones a tomar por las entidades gubernamentales serán más precisas.
 - o Rapidez de carga y proceso de los mecanismos analíticos.
 - o Rapidez de carga y proceso de la herramienta de visualización.
 - o Número de insights accionables a partir del perfilamiento. Mientras más insights se obtengan del perfilamiento, más iniciativas se pueden crear a favor de reducir la brecha digital.
 - o Nivel de automatización de la solución. Mientras más cómoda sea la experiencia para el stakeholder, mejor.
 - o Reducción del tiempo de implementación de la solución en los sistemas del stakeholder.
- Atractivos
 - o Mecanismos analíticos de la solución involucran Inteligencia Artificial Generativa.
 - o Paleta de colores corporativa del stakeholder utilizada en la solución.
 - o Horas de capacitación para los usuarios de negocio que consumirán la solución.
 - o Alertas en tiempo real de insights accionables.
 - o Personalización de la solución después de un año de prueba.

2. ¿Quiénes o qué áreas deberán participar en la solución analítica?

Participarán en la solución 4 recursos, incluyendo un líder de proyecto, un científico de datos, un analista de datos senior y un analista en visualización de datos. Además, se necesitará trabajar con las áreas de tecnología y el área de analítica del stakeholder para analizar la factibilidad de implementar la solución en sus sistemas. También, deberá participar algún miembro de las entidades gubernamentales para indicar la factibilidad de los insights obtenidos a partir del perfilamiento de la población objetivo.

3. ¿Cómo la implementación de estos requerimientos impactaría en los objetivos del proyecto?

Requerimientos a incrementar

Nombre y Descripción	Escala de viabilidad de desarrollo e implementación (1-10)	Escala de impacto en el objetivo del proyecto (1-10)	Score total
El proyecto debe ayudar a CNC a generar un entendimiento del proceso	10	10	10
En la solución del proyecto se deben utilizar variables explicativas	10	10	10
El proyecto debe generar al menos dos mecanismos analíticos	10	10	10
Se debe generar una herramienta de visualización de los modelos de la	10	10	10
La herramienta de visualización de los modelos de la solución se debe	9	6	7.5
La solución del proyecto se debe disponibilizar como un servicio en la	7	6	6.5
El desarrollo de los mecanismos analíticos del proyecto se debe	10	6	8
Los insumos de datos del proyecto se deben alojar en los servidores	7	8	7.5
Porcentaje de la población objetivo perfilada. Mientras más personas	7	10	8.5
Rapidez de carga y proceso de los mecanismos analíticos.	8	8	8
Rapidez de carga y proceso de la herramienta de visualización.	10	8	9
Número de insights accionables a partir del perfilamiento. Mientras	8	10	9
Nivel de automatización de la solución. Mientras más cómoda sea la	8	10	9
Reducción del tiempo de implementación de la solución en los	7	9	8
Mecanismos analíticos de la solución involucran Inteligencia Artificial	6	8	7
Paleta de colores corporativa del stakeholder utilizada en la solución.	10	6	8
Horas de capacitación para los usuarios de negocio que consumirán la	10	10	10
Alertas en tiempo real de insights accionables.	7	6	6.5
Personalización de la solución después de un año de prueba.	10	10	10

Requerimientos para incrementar

Nombre y Descripción	Escala de viabilidad en la reducción en desarrollo e implementación (1-10)	Escala de impacto en el objetivo del proyecto (1-10)	Score total
La herramienta de visualización debe tener varias ventanas con la mayor información posible.	10	8	9
Los mecanismos analíticos desarrollados deben dividir a la población objetivo en muchos subgrupos (al menos 5).	8	8	8
El acceso a la solución analítica debe ser gestionado por el stakeholder. (Puesta en producción y disponibilización).	6	8	7
La solución analítica solo se puede usar para los datos otorgados por el stakeholder y no por datos nuevos.	10	10	10
La solución debe utilizar proveedores que involucren costos extra.	10	10	10

4. Prioriza los requerimientos generando un orden o decisión de aquellos que van a ser incorporados al proyecto de acuerdo con el criterio de tu equipo. Considera que la escala de desarrollo en la implementación o su contrapartida puede también involucrar aspectos tecnológicos, monetarios o de otra índole que se debieran tener en cuenta en el desarrollo. Estos requerimientos se conocen también como los key-Drivers del proyecto. Ten presente que regularmente, estas evaluaciones se validan con el equipo y los SH, y se ajustan de acuerdo a ellos.

Key-Drivers del proyecto

Priorización	Nombre y Descripción	Score total
1	El proyecto debe ayudar a CNC a generar un entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia.	10
2	En la solución del proyecto se deben utilizar variables explicativas como penetración y acceso.	10
3	El proyecto debe generar al menos dos mecanismos analíticos diferentes apoyados en modelos tipo White, Gray o Black-Box Model.	10
4	Se debe generar una herramienta de visualización de los modelos de la solución.	10

Priorización	Nombre y Descripción	Score total
5	Horas de capacitación para los usuarios de negocio que consumirán la solución.	10
6	Personalización de la solución después de un año de prueba.	10
7	(Reducir) La solución analítica solo se puede usar para los datos otorgados por el stakeholder y no por datos nuevos.	10
8	(Reducir) La solución debe utilizar proveedores que involucren costos extra.	10
9	(Aumentar) Número de insights accionables a partir del perfilamiento. Mientras más insights se obtengan del perfilamiento, más iniciativas se pueden crear a favor de reducir la brecha digital.	9
10	(Aumentar) Rapidez de carga y proceso de la herramienta de visualización.	9
11	(Aumentar) Nivel de automatización de la solución. Mientras más cómoda sea la experiencia para el stakeholder, mejor.	9
12	(Reducir) La herramienta de visualización debe tener varias ventanas con la mayor información posible.	9

5. Valida, a partir de estos requerimientos, las métricas iniciales de éxito y fracaso identificadas en la GUÍA 1. Es decir, a partir del problema de negocio y la métrica de negocio inicial establecida en la GUÍA 1, descompón esa métrica en subpartes asignables a cada requerimiento seleccionado en el paso anterior. Cuando se cuenta con ellos, este paso debe ser negociado y revisado junto con los SH, para que las métricas sean posteriormente conocidas, medibles y trazables.

La métrica de negocio inicial establecida considera que:

El stakeholder obtiene un perfilamiento más preciso de la población de interés (individuos con baja apropiación digital). Se espera que al menos el 80% de los individuos sean perfilados adecuadamente, ya que corresponderían a la mayoría de la población total.

Esto se puede descomponer y asociar a los requerimientos de la siguiente manera:

Requerimiento	Subparte de la métrica de negocio
El proyecto debe ayudar a CNC a generar un entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia.	Tener un entendimiento del proceso y dinámica de Adopción Digital ayudará a entender quienes forman parte de la población objetivo.
En la solución del proyecto se deben utilizar variables explicativas como penetración y acceso.	Estas variables serán de ayuda para perfilamiento.
El proyecto debe generar al menos dos mecanismos analíticos diferentes apoyados en modelos tipo White, Gray o Black-Box Model.	Los mecanismos analíticos son necesarios para generar un perfilamiento matemáticamente preciso.
Se debe generar una herramienta de visualización de los modelos de la solución.	Para que el stakeholder pueda entender el perfilamiento necesita consumirlo de forma intuitiva.
Horas de capacitación para los usuarios de negocio que consumirán la solución.	Para asegurarnos que el stakeholder use la solución de forma adecuada se lo debe capacitar.
Personalización de la solución después de un año de prueba.	Con este requerimiento nos aseguramos de que una vez madurado el análisis el stakeholder siga teniendo un perfilamiento de al menos 80%.
(Reducir) La solución analítica solo se puede usar para los datos otorgados por el stakeholder y no por datos nuevos.	Para que la solución siga funcionando en el tiempo tiene que estar habilitada para recibir datos nuevos, o el porcentaje de individuos perfilados adecuadamente decaerá.
(Reducir) La solución debe utilizar proveedores que involucren costos extra.	El stakeholder espera una solución autónoma que no dependa de proveedores externos.
(Aumentar) Número de insights accionables a partir del perfilamiento. Mientras más insights se obtengan del perfilamiento, más iniciativas se pueden crear a favor de reducir la brecha digital.	El número de insights accionables se correlaciona con la cantidad de individuos perfilados adecuadamente.
(Aumentar) Rapidez de carga y proceso de la herramienta de visualización.	Un procesamiento de datos eficiente agiliza el entendimiento de las características de los individuos.
(Aumentar) Nivel de automatización de la solución. Mientras más cómoda sea la experiencia para el stakeholder, mejor.	El stakeholder requiere autonomía para poder perfilar a los individuos con los resultados de la solución.
(Reducir) La herramienta de visualización debe tener varias ventanas con la mayor información posible.	Para que el entendimiento de los resultados sea adecuado se debe mostrar solo la información necesaria.

Fase d) Los recursos

Con base en los requerimientos establecidos y su jerarquización, establece una relación inicial de los recursos requeridos para desarrollarlos. Algunos de estos recursos son:

- Personal y perfiles requeridos.
Dados los requerimientos establecidos, se requerirán 4 recursos de personas con los siguientes perfiles y funciones:
 - Líder de proyecto/Gerente de proyecto
Será el recurso encargado de alinear a los demás recursos con el objetivo del proyecto. También, administrará la carga de trabajo, revisará avances y retroalimentará al equipo en base a la retroalimentación otorgada por el stakeholder. En otras palabras, es el recurso que tomará la responsabilidad de materializar el entregable del proyecto ajustado a las restricciones y a las necesidades del negocio.
 - Científico de datos

Será el encargado de desarrollar los dos mecanismos analíticos requeridos. Además, trabajará alineado a que la solución planteada genere insights accionables a partir del perfilamiento creado con la información otorgada, automatizando lo más posible los procesos y asegurando que el entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia sea adecuado.

3. Analista de datos senior

Trabjará transversalmente con el Científico de datos y el Analista en visualización de datos para analizar que la información otorgada por el stakeholder sea de utilidad para la solución analítica, tanto en el desarrollo de los mecanismos de analítica, generación de insights, así como en las visualizaciones correspondientes. En otras palabras, este recurso brindará apoyo operativo en el entendimiento y la transformación de los datos a utilizar.

4. Analista en visualización de datos

Será el responsable de disponibilizar los resultados para los usuarios finales a través de herramientas de visualización de datos, siempre priorizando que el procesamiento de los datos sea eficiente, y que la información presentada no sea excesiva, sino que ayude a generar un entendimiento del proceso y dinámica de Adopción Digital (AD) en Colombia utilizando variables explicativas como penetración y acceso.

b. Tiempo de implementación.

El tiempo de implementación estimado es de 2 meses, considerando el primer día para definir y entender la problemática del negocio planteada en estos documentos del proyecto. Luego, una semana para la recolección de datos facilitada por el stakeholder. En la segunda semana se propone analizar los datos y realizar la segmentación de la población para que en las próximas dos semanas se pueda mejorar el análisis al entender y perfilar a las poblaciones objetivos de donde se obtendrán los insights de interés para el stakeholder. Finalmente, el último mes se dedicará para desarrollar las visualizaciones necesarias para automatizar lo más posible la obtención de insights para el accionar de las entidades gubernamentales, también diseñando un sistema de seguimiento para monitorear la efectividad de las intervenciones y medir constantemente la evolución de las métricas de adopción digital.

c. Otros recursos tales como tecnológicos.

Además de los recursos de tiempo y personal, se requerirá de recursos tecnológicos para el desarrollo de la solución analítica. Por un lado, se necesitarán 4 equipos de computación con la capacidad necesaria para elaborar tareas de ofimática. El desarrollo y modelamiento de los mecanismos analíticos solicitados se realizarán en proveedores de procesamiento en la nube, mediante herramientas de Software as a Service (SaaS), con un requerimiento de procesamiento de alrededor de 8 cores de cpu, 32GB de RAM y 100GB de almacenamiento para cada miembro del personal. Además, se necesitarán de licencias en las herramientas de desarrollo de los mecanismos de analítica y las visualizaciones. Por un lado, se utilizará Jupyter Notebook y para las visualizaciones se utilizará PowerBi.

d. Presupuesto de la inversión. (Supongan un aproximado de acuerdo con la información que cuentan).

Considerando los recursos humanos y tecnológicos que se requieren para el proyecto se estima un presupuesto total de \$42,393,000 COP, el cual se desglosa de la siguiente manera:

Recurso	Costo
Salarios Líder de proyecto y analistas (2 meses)	\$40,000,00 COP
Alquiler equipos de computación	\$1,700,000 COP
Créditos herramientas de procesamiento en la nube (SaaS)	\$693,000 COP

Fase e) El planteamiento inicial del Problema de Analytics

Escribe una descripción articulada de problema a abordar empleando desde los requerimientos del primer punto hasta los resultados del punto anterior. Esta descripción deberá considerar los siguientes aspectos:

- Objetivos a abordar con base en los requerimientos validados en la solución
- Presupuesto de recursos, tales como económico, humano, etc.
- Tiempo disponible de solución.
- Resultado o producto esperado.

Al tener este planteamiento inicial debes asegurarte de que existe coherencia con el planteamiento de problema; es decir los requerimientos y objetivos derivados están ligados con el problema de negocio.

Problema de Analytics:

Con el objetivo de ayudar a CNC a generar un entendimiento del proceso y dinámica de la adopción digital (AD) en Colombia y así poder brindar insights accionables para las entidades gubernamentales, se busca realizar un perfilamiento de la población objetivo donde la adopción digital es escasa, el cual se disponibilizará mediante un Dashboard. La solución se realizará con un equipo conformado por 4

integrantes, incluyendo un líder de proyecto, un científico de datos, un analista de datos senior y un analista en visualización de datos, que, mediante herramientas de lenguaje de Python y recursos computacionales de la nube, diseñarán al menos dos mecanismos de analítica y una herramienta de visualización en un lapso de dos meses con un presupuesto total de \$43,393,000 COP.

Parte II. Establecimiento de las fuentes y los objetivos de análisis de los datos.

Fase a) Establecer las fuentes.

1. En la GUÍA 1 respondieron las preguntas: ¿Consideran que los datos potenciales a emplear existen o se pueden generar? ¿Dónde se encuentran estos o cómo se podrían generar? acorde el planteamiento analítico y avance al momento ¿identificas que se requieren más fuentes de datos que las planteadas en la GUÍA 1? En caso afirmativo, describe cuáles serían estas fuentes de información. En la guía 1 se definió que los datos a emplear ya existen. Se cuenta con una base de datos de más de 20 mil encuestas realizadas a casas hogar (2016-2023) realizadas por el CNC. Además, estos datos se complementan con información como la tasa de penetración de internet, con reportes de proyecciones del Departamento Administrativo nacional de Estadística de Colombia (DANE) y con reportes de acceso obtenidos del Instituto Colombiano para la Evaluación de la Educación (ICFES).
2. ¿Los datos identificados son coherentes para abordar el problema de negocio planteado o se tienen que replantear?
Los datos disponibles son suficientes y relevantes para abordar el objetivo del proyecto. La base de datos de las encuestas levantadas por el CNC y los datos del Saber 11 de Colombia proporciona una gran cantidad de información tanto de los hogares como de los niños, lo cual nos permite perfilar a las poblaciones y entender su nivel de apropiación digital. Al complementar esta base de datos con estadísticas de conectividad y acceso a internet (DANE, ICFES), podemos construir un panorama integral.
3. ¿Las fuentes adicionales de información requieren el desarrollo y aplicación de algún instrumento (ej., encuesta)? Si sí, describe. Nota: considera que requerirás tiempo para obtenerlos.
No consideramos necesario desarrollar y aplicar un nuevo instrumento, como una encuesta, en esta etapa, ya que contamos con datos adecuados para los objetivos actuales del análisis. En caso de que en futuras entregas se identifique la necesidad de información adicional, esta debería ser de carácter histórico para alinearse con el periodo de análisis (2012 - 2023), permitiendo una integración coherente con los datos ya disponibles.
4. Para los datos que existen y los que se tengan que generar, indica qué tipo de datos son (ej., series de tiempo, datos no estructurados, datos panel, entre otros).
 - **Encuesta a los hogares realizada por el CNC:** Estos datos corresponden a muestras transversales, ya que se recogen observaciones de hogares distintos en cada periodo (2016-2023).
 - **Datos del Saber 11:** Estos datos corresponden a un solo corte por cada grupo de individuos y no están organizados en panel. Son observaciones de tipo transversal, recolectadas para cada cohorte que toma el examen en un año específico.
 - **Otras fuentes (reportes):** Los reportes de otras instituciones (DANE, ICFES) corresponden a datos no estructurados. Estos incluyen proyecciones de penetración de internet, estadísticas nacionales, y otros informes que pueden contener texto, tablas, o gráficos sin una estructura relacional específica.

Esta variedad en los tipos de datos permite realizar un análisis integral al combinar información estructurada y no estructurada sobre el contexto socioeconómico y la apropiación digital de los hogares.

5. Con base en los requerimientos establecidos y tu respuesta anterior, especifica si es necesaria alguna transformación que genere datos útiles para el análisis. Ejemplo, si consideras datos basados en observaciones humanas, cuestionarios, opiniones de redes sociales.
 - **Encuesta a los hogares realizada por el CNC:** Los datos de esta encuesta están organizados con registros individuales dispuestos horizontalmente en las columnas. Dado que el objeto de análisis es el hogar o la familia, será necesario realizar un proceso de *feature engineering* para transformar estas observaciones individuales en atributos representativos de cada hogar, mejorando así la calidad y pertinencia de los datos para el análisis.
 - **Datos del Saber 11:** Aunque estos datos están estructurados con un propósito distinto, es posible transformarlos para nuestro objetivo. Este proceso implicará una preparación de datos específica para extraer y organizar la información relevante que nos permita analizar la apropiación digital y el perfil sociodemográfico de los hogares de los estudiantes.
 - **Reportes de otras fuentes (DANE, ICFES):** Dado que los reportes están en un formato no estructurado, es necesario realizar un procesamiento de la información para integrarla en el análisis. Esto incluye extraer, limpiar y posiblemente estandarizar las métricas relevantes (como estadísticas de penetración de internet y proyecciones poblacionales) para que se puedan relacionar adecuadamente con los otros datos.
6. Si es posible, identifica qué departamentos o áreas de la organización están involucradas con las fuentes de datos que dan respuesta a los requerimientos antes establecidos.

Aunque las fuentes de datos provienen de entidades externas, el proceso de recolección y gestión de esta información es interno. Por lo tanto, las áreas clave involucradas en el ciclo de vida de los datos dentro de la organización son:

1. **Área de Recolección de Información:** Responsable de gestionar la obtención y recopilación de los datos externos requeridos.
2. **Gobierno de Datos:** Encargada de asegurar la calidad, integridad y conformidad de los datos, garantizando su adecuación a los estándares de la organización.
3. **Área de Tecnologías:** Responsable de proporcionar la infraestructura y herramientas necesarias para el almacenamiento, procesamiento y análisis de los datos.

Estas áreas colaboran para facilitar un flujo de datos eficiente y asegurar que la información sea confiable y utilizable a lo largo del análisis.

Fase b) Establecer las fuentes.

1. Acorde al planteamiento de negocio y requerimientos previamente establecidos, ¿cuál(es) sería el objetivo de análisis de los datos para responder a este problema? Redáctalo(s).

- Caracterizar el nivel de adopción digital en diferentes regiones y grupos demográficos, proporcionando una visión detallada del estado actual y las disparidades demográficas en el acceso digital.
- Determinar el impacto de variables sociodemográficas y de conectividad sobre la adopción digital, identificando cuáles de estas variables contribuyen de manera significativa a la adopción o limitación de la digitalización en el país.
- Preparar los datos para modelar el proceso de adopción digital mediante enfoques analíticos tipo White, Gray o Black Box-Model, lo cual servirá como base para construir los mecanismos que aporten a los stakeholders un entendimiento claro y fundamentado del fenómeno.

2. ¿Qué nivel de transformación de los datos generaría una respuesta a tu problema de negocio?

Para que los modelos de analítica de datos puedan ingerir los datos presentes en las encuestas y otras bases de datos dentro hace falta un tratamiento profundo a los datos. En primer lugar, hace falta estandarizar las fuentes de información para simplificar el problema en una sola base de datos más otras variables fijas y temporales. Segundo, se debe transformar información cualitativa en cuantitativa de tal forma que pueda ser consumida por los modelos de detección de patrones. Finalmente, se debe versionar estos cambios y tratamientos para que el proceso sea auditable y gobernable para su sostenibilidad en el tiempo.

3. ¿Qué modelos y técnicas podrían dar solución al problema?

Para abordar el problema de la adopción digital en Colombia, se pueden implementar técnicas no supervisadas que faciliten la reducción de dimensionalidad y la identificación de patrones en los datos. El Análisis de Componentes Principales (PCA) permitirá simplificar la complejidad del conjunto de datos al condensar la información en un número reducido de variables, lo que facilitará el análisis posterior.

Además, se puede aplicar la clusterización, utilizando algoritmos como K-means o DBSCAN, para segmentar la población según sus características de adopción digital. Esto permitirá identificar diferentes raíces de la no adopción y comprender mejor las características y necesidades específicas de cada grupo.

4. Posteriormente, ¿qué actividades implementarías para dar respuesta al problema de negocio?

1. **Estructuración de Datos:** Se comenzaría por organizar y estructurar los datos provenientes de las diferentes fuentes, como la Encuesta a los hogares realizada por el CNC, los datos del Saber 11 y los reportes del DANE e ICFES. Esta etapa es crucial para garantizar que la información sea coherente y fácil de analizar.
2. **Estandarización de Datos:** Posteriormente, se realizaría un proceso de estandarización para unificar los formatos y estructuras de los datos, formando un conjunto de datos homogéneo que permita un análisis más efectivo. Esto incluiría la limpieza de datos para eliminar inconsistencias y la creación de nuevas variables que faciliten el análisis.
3. **Análisis Exploratorio y Modelado:** Una vez que los datos estén preparados, se procederá con un análisis exploratorio para identificar patrones y correlaciones relevantes. Luego, se implementarán técnicas no supervisadas, como la reducción de dimensionalidad y la clusterización, para profundizar en la dinámica de la adopción digital y determinar las raíces de la no adopción.

4. Desarrollo de Modelos Analíticos: Finalmente, se desarrollarán los modelos analíticos tipo White, Gray o Black Box para proporcionar a los stakeholders un entendimiento claro y fundamentado del proceso de adopción digital en Colombia. Estos modelos ayudarán a guiar la toma de decisiones y el diseño de estrategias efectivas para mejorar la adopción digital en el país.

Fase c) Determinar criterios de éxito en el análisis

- 1) Describe los métodos para la evaluación de cierto modelo que selecciones, tal como la exactitud, precisión y desempeño.

Para determinar el éxito en el análisis de la adopción digital en Colombia técnicas no supervisadas como PCA y clustering, se deben considerar criterios que evalúen tanto la calidad del modelo con su aplicabilidad e impacto en la comprensión del objetivo.

Calidad del modelo:

- Varianza explicada (PCA): Este es un criterio clave, la cantidad de varianza explicada por los componentes principales seleccionados. El propósito es identificar un número reducido de componentes que tengan un alto porcentaje de la varianza principal.
- Separación de clusters: Estos deben ser bien definidos y diferenciados entre sí. Esto se puede evaluar con métricas como el índice de silueta o el coeficiente de Calinski-Harabasz.
- Interpretabilidad de clusters: Los clusters deben ser interpretables en el contexto de la adopción digital. Cada cluster debe representar un grupo con características distintas y significativas en cuanto a los diferentes factores determinados.

Aplicabilidad e impacto:

- Identificación de patrones: El análisis debe revelar patrones relevantes en los datos, por ejemplo, grupos de baja adopción, características comunes entre los no adoptantes o factores asociados a una mayor adopción.
- Accionabilidad: Los resultados del análisis realizado deben ser útiles para la toma de decisiones. Los clusters identificados deben permitir la segmentación de la población para el desarrollo de estrategias específicas que promuevan la adopción digital.
- Impacto en políticas públicas: Idealmente, el análisis debe contribuir a la formulación de políticas públicas que tengan un mayor impacto y sean más efectivas para disminuir la brecha digital en Colombia.

- 2) Define las métricas *benchmark* para evaluar el éxito, siendo claro en proveer métricas específicas.

Métricas benchmark

Para establecer estas métricas, se pueden emplear estudios previos o datos de referencia sobre la adopción digital en Colombia.

- Estudios previos:

- Revisión de literatura: Investigar estudios previos que hayan utilizado modelos de PCA o clustering para analizar la adopción digital en Colombia o países con características similares. Identificar las métricas empleadas en dichos estudios para tener un punto de referencia.
- Bases de datos académicas: Consultar bases de datos para encontrar artículos relevantes y extraer información sobre las métricas utilizadas.
- Informes institucionales: Validar informes de entidades como el MinTIC, el DANE que contengan datos sobre esta problemática y las métricas utilizadas para el análisis.

Al establecer métricas de benchmark claras y específicas, se podrá evaluar de manera objetiva el éxito del análisis realizado, como también poder determinar si el modelo cumple con los criterios de calidad, aplicabilidad e impactos deseados.

- 3) Define las métricas subjetivas, si las tienes, de la mejor manera, así como determina aquella que represente un éxito.

Una métrica subjetiva puede ser la interpretabilidad de los clusters. Ya que permiten un análisis cualitativo, que se podría considerar exitoso si facilita la comprensión de las diferentes razones de la adopción y que sean útiles para diseñar estrategias e intervenciones específicas.

Fase d) Tipos de problemas de Analytics

1. Acorde los objetivos de análisis de los datos y criterios de éxitos en el análisis, ¿cuál es o son los tipos de enfoques y técnicas que identificas son candidatos para dar respuesta al problema de negocio?

Comprender los antecedentes

- Análisis descriptivo: Caracterizar la población con baja adopción digital en términos de las variables demográficas, socioeconómicas, geográficas, de acceso a tecnología y uso de internet.
- Visualización de datos: Crear visualizaciones (mapas, histogramas, gráficos, etc.) para mostrar la distribución de la adopción digital en Colombia.

Segmentación de datos/detección de anomalías (Modelo no supervisado)

- Kmeans - DBSCAN: Segmentar la población con baja adopción digital en grupos con características y necesidades similares.
- PCA: Reducir la dimensionalidad de los datos para facilitar la visualización e interpretación de los clusters. Como también la identificación de variables que mas contribuyen a la varianza.

2. ¿De qué manera estas técnicas dan respuesta a los objetivos de análisis antes establecidos?

- El análisis descriptivo y visualización permitirán entender la situación actual de la problemática en estudio, identificar las áreas con mayor necesidad de intervención y comunicar los hallazgos de forma clara y concisa.
- Clustering: Facilitará la comprensión de las diferentes características de la población y permitirá diseñar estrategias e intervenciones diferenciadas para cada segmento.
- PCA: Ayudará a identificar las variables más relevantes para explicar la varianza en la adopción digital, lo que puede orientar la toma de decisiones y la asignación de recursos.

3. ¿Cómo se alinearían los resultados de la aplicación de estas técnicas, ante el problema de negocio?

- El análisis permitirá a CNC generar un entendimiento profundo de la AD en Colombia.
- Los insights generados serán accionables, ya que permitirán a las entidades gubernamentales diseñar políticas públicas e intervenciones más efectivas para promover la inclusión digital.
- El dashboard facilitará la comunicación de los resultados a las entidades gubernamentales y otros stakeholders.
- El proyecto se alinea con el presupuesto y el tiempo disponible.

4. ¿Qué acciones implementarías en cada caso de la aplicación de técnicas para dar respuesta al problema de negocio?, ¿se requeriría la aplicación de otro nivel más de técnicas? (lo que comentamos del sentido de generar patrones de técnicas a los datos) Si sí, describir.

Comprender los antecedentes:

- Obtener datos de fuentes confiables (DANE, MinTIC).
- Realizar análisis descriptivo de variables demográficas, socioeconómicas, geográficas, de acceso a tecnología y uso de internet.
- Crear visualizaciones informativas (mapas, histogramas, gráficos de dispersión) para el dashboard.

Segmentación de datos/detección de anomalías (Modelo no supervisado):

- Aplicar PCA para reducir la dimensionalidad.
- Implementar K-means o DBSCAN para segmentar la población.
- Evaluar la calidad de los clusters (índice de silueta, etc.).
- Interpretar y describir las características de cada cluster.
- Visualizar los clusters en el dashboard.

Fase e) Planteamiento Problema Analytics

1. Capturar y mantener la descripción del problema a resolver.

El problema por resolver se centra en la necesidad del Centro Nacional de Consultoría (CNC) de comprender el proceso y la dinámica de la adopción digital en Colombia. Esto con el fin de generar insights accionables que permitan a las entidades gubernamentales diseñar políticas públicas e intervenciones más efectivas para reducir la brecha digital y promover la inclusión digital en el país.

2. Ayudar a especificar todas las preguntas, hitos de negocio de la manera más precisa posible.

Pregunta principal: ¿Cómo se puede caracterizar y segmentar la población colombiana con baja adopción digital para generar insights que faciliten el diseño de políticas públicas e intervenciones que promuevan la inclusión digital?

Hitos de negocio:

- Identificar las características de la población con baja adopción digital.
- Segmentar la población con baja adopción digital en grupos con necesidades similares.
- Identificar las variables que más influyen en la adopción digital.

d. Generar un dashboard con insights accionables para las entidades gubernamentales.

3. Alinearse con los requerimientos de negocio.

El planteamiento del problema se alinea con los requerimientos del negocio de la siguiente manera:

- **Enfocarse en la comprensión de la Adopción Digital:** El análisis se centra en la caracterización y segmentación de la población con baja adopción digital, lo que responde a la necesidad del CNC de entender este fenómeno.
 - **Generar insights accionables:** El análisis busca generar información útil para el diseño de políticas públicas e intervenciones, lo que se alinea con el objetivo de brindar apoyo a las entidades gubernamentales.
 - **Utilizar los datos disponibles:** El análisis se basa en datos existentes del CNC, DANE e ICFES, lo que optimiza los recursos y el tiempo.
 - **Desarrollar un dashboard:** Se busca crear una herramienta de visualización que facilite la comunicación de los resultados a las entidades gubernamentales.
4. Permitir derivar los beneficios esperados en términos de negocio.
- **Mejorar la toma de decisiones:** Al brindar un entendimiento más profundo de la adopción digital, el análisis permitirá a las entidades gubernamentales tomar decisiones más informadas sobre políticas públicas e intervenciones.
 - **Optimizar la asignación de recursos:** Al identificar los grupos con mayor necesidad y las variables que más influyen en la adopción digital, se podrán optimizar los recursos destinados a la promoción de la inclusión digital.
 - **Aumentar la eficiencia de las intervenciones:** Al segmentar la población y comprender las diferentes necesidades, se podrán diseñar intervenciones más específicas y efectivas.

Parte III. Estado del arte: referencias de investigaciones, aplicaciones en el área:

Estado del Arte

El presente estado del arte analiza investigaciones y estudios recientes relacionados con la inclusión y la brecha digitales en Colombia y América Latina. A continuación, se describen las contribuciones clave de estos trabajos, destacando cómo se alinean con el planteamiento del problema, la metodología y los objetivos del proyecto, así como sus hallazgos principales.

Inclusión Digital y Apropiación de las TIC

Salinas y de Benito (2020) exploran el impacto de las políticas educativas y de infraestructura en la inclusión digital en América Latina. Su estudio revela que la simple provisión de tecnología no garantiza la inclusión digital efectiva. A través de un análisis de las políticas implementadas, identifican que el acceso y la conectividad deben ir acompañados de la capacitación en competencias digitales para asegurar que los usuarios puedan aprovechar plenamente las TIC. Además, el estudio destaca que el enfoque integral, que incluye tanto infraestructura como formación, es clave para superar las barreras tecnológicas en comunidades marginadas. Este enfoque se refleja en el proyecto actual, que busca no solo mejorar la conectividad, sino también fomentar la alfabetización digital y la adopción significativa de tecnologías digitales, especialmente en zonas rurales (Salinas & de Benito, 2020).

Brecha Digital en la Educación Rural Colombiana

La revisión sistemática realizada por Moreno Lizarazo (2023) se centra en la brecha digital en las áreas rurales de Colombia, subrayando que la falta de acceso a internet y a recursos tecnológicos continúa siendo una barrera significativa para la equidad educativa. Mediante la metodología PRISMA que el autor plantea, se identificaron factores críticos como la infraestructura deficiente y la escasa inversión en tecnologías para las zonas rurales. Los resultados sugieren que, aunque se han implementado políticas para reducir la brecha digital, estas han sido insuficientes en áreas geográficamente aisladas. Este estudio es particularmente relevante para el proyecto, ya que subraya la necesidad de que todo proyecto que se implemente tenga un foco de intervención focalizada que aborde la conectividad y la infraestructura, así como programas de capacitación en habilidades digitales para poblaciones rurales.

Transformación Digital en Colombia: Desafíos y Oportunidades

El informe de la OCDE (2019) proporciona una visión amplia sobre la transformación digital en Colombia, destacando los avances y desafíos en la adopción de tecnologías digitales en sectores clave del país. Según el informe, aunque Colombia ha mejorado sus tasas de conectividad y

adopción tecnológica, persisten desafíos importantes relacionados con **la infraestructura en zonas rurales y la alta disparidad regional**. El estudio enfatiza la necesidad de políticas coherentes que promuevan la digitalización y faciliten la inclusión digital, especialmente en sectores como la educación y las pequeñas empresas. Las recomendaciones de la OCDE incluyen mejorar la infraestructura de banda ancha, reducir barreras de costos para la adquisición de tecnología y fortalecer la capacitación en competencias digitales. Estos hallazgos son congruentes con el proyecto actual, que busca apoyar a las entidades gubernamentales (decisores) con el perfilamiento para diseñar políticas públicas basadas en evidencia y que faciliten el acceso a las TIC, la capacitación en habilidades digitales y la reducción de la brecha digital en regiones menos favorecidas de Colombia (OCDE, 2019).

Encuesta Nacional de Calidad de Vida 2020

La Encuesta Nacional de Calidad de Vida (2020), realizada por el Departamento Administrativo Nacional de Estadística (DANE), ofrece datos actualizados sobre el acceso a servicios tecnológicos y conectividad en Colombia. Según la encuesta, existe una disparidad significativa en el acceso a tecnologías entre las áreas urbanas y rurales, con una menor proporción de hogares rurales que cuentan con acceso a internet. Esta falta de conectividad limita las oportunidades educativas y económicas para las comunidades rurales. Los datos de la ECV refuerzan la necesidad de políticas públicas que promuevan la equidad en el acceso a tecnologías digitales, algo que el proyecto actual también aborda al proponer iniciativas específicas para mejorar la infraestructura tecnológica en regiones rurales (DANE, 2020).

Resumen y aportes

Los estudios revisados coinciden en que la brecha digital en Colombia y América Latina sigue siendo un desafío multifacético que requiere soluciones integrales. Las políticas actuales han logrado avances en mejorar la conectividad, pero persisten barreras significativas, especialmente en áreas rurales. Los estudios subrayan que el acceso a infraestructura tecnológica debe ir acompañado de programas de capacitación que permitan a las comunidades hacer un uso significativo de las TIC, fomentando así el desarrollo personal, educativo y económico.

De igual manera, los estudios realizados aportan varias similitudes y elementos clave que respaldan la construcción de la solución planteada. En primer lugar, el estudio de Salinas y de Benito (2020) destaca la importancia de no solo proporcionar infraestructura tecnológica, sino también de comprender el contexto y las necesidades de las comunidades para fomentar una adopción digital efectiva. Este enfoque se refleja en la solución planteada, que busca realizar un **perfilamiento detallado de las poblaciones con baja adopción digital, permitiendo así identificar barreras específicas y diseñar intervenciones más focalizadas**.

Asimismo, la revisión sistemática de Moreno Lizarazo (2023) sobre la brecha digital en las áreas rurales de Colombia subraya la necesidad de intervenciones específicas y dirigidas para cerrar las desigualdades tecnológicas, algo que **se alinea directamente con el objetivo de la solución propuesta de generar insights accionables para entidades gubernamentales**. La literatura también destaca la efectividad de utilizar datos y análisis para informar políticas públicas, tal como lo indica el informe de la OCDE (2019), que promueve el uso de herramientas analíticas y tecnológicas para abordar desafíos nacionales, incluyendo la adopción digital. Esto resuena con la metodología planteada en la solución analítica, que emplea técnicas de *machine learning* y otras tecnologías emergentes para desarrollar mecanismos de análisis y visualización que faciliten la toma de decisiones.

En conjunto, la literatura no solo valida la importancia de un enfoque analítico en la adopción digital, sino que también proporciona un marco conceptual sólido que apoya el diseño y la implementación de la solución planteada.

Referencias

- DANE. (2020). *Encuesta Nacional de Calidad de Vida 2020*. Bogotá: Departamento Administrativo Nacional de Estadística.
- Moreno Lizarazo, C. (2023). *La brecha digital en la educación rural colombiana desde una revisión sistemática*. *Dialéctica*, 21, 141-145.
- OCDE. (2019). *OECD Reviews of Digital Transformation: Going Digital in Colombia*. París: OECD Publishing.
- Salinas, J., & de Benito, B. (2020). *Competencia digital y apropiación de las TIC: Claves para la inclusión digital*. *Campus Virtuales*, 9(2), 99-111.