Manual de Usuario para el Monitor de Inclusión Digital Colombia

1. Introducción

El Monitor de Inclusión Digital Colombia es una herramienta avanzada diseñada para el análisis de la adopción digital en Colombia. Esta plataforma, basada en los datos proporcionados por la Comisión Nacional de Telecomunicaciones (CNC) en 2023, permite identificar patrones de comportamiento digital a partir de la segmentación de la población. Utilizando metodologías estadísticas como el FAMD (Factor Analysis of Mixed Data) y el K-Means clustering, el sistema categoriza a los usuarios en diferentes segmentos, permitiendo analizar los niveles de adopción digital en relación con variables como el acceso a la tecnología, las habilidades digitales y el entorno socioeconómico. La plataforma incluye visualizaciones interactivas a través de mapas y gráficos estadísticos en tiempo real, y proporciona recomendaciones basadas en análisis automáticos, todo con el fin de facilitar la toma de decisiones para la creación de políticas públicas de inclusión digital.

Funcionalidades principales:

- **Carga de Datos Personalizada:** Capacidad de incorporar nuevos conjuntos de datos, lo que permite regenerar los análisis y visualizaciones correspondientes.
- **Análisis de Segmentación:** Identificación de segmentos poblacionales basados en el comportamiento digital mediante FAMD y K-Means, utilizando más de 15 variables relacionadas con el acceso y uso de tecnología.
- **Insights con Inteligencia Artificial (IA):** Análisis automático de los segmentos y generación de recomendaciones para políticas públicas utilizando IA.
- **Visualización Interactiva:** Provisión de un dashboard web que incluye mapas geográficos interactivos y gráficos estadísticos actualizados en tiempo real.

El Monitor de Inclusión Digital Colombia ofrece diversas ventajas que facilitan el análisis y la toma de decisiones en cuanto a la adopción digital en el país. Una de las principales ventajas es la capacidad de segmentar la población de manera precisa, basándose en un amplio conjunto de variables relacionadas con el acceso a la tecnología, las habilidades digitales y otros factores demográficos. Este análisis segmentado permite diseñar políticas públicas más efectivas, adaptadas a las necesidades específicas de cada grupo de población.

Además, el sistema integra inteligencia artificial para proporcionar recomendaciones automáticas, lo cual es una herramienta poderosa para los responsables de la toma de decisiones, ya que les permite identificar oportunidades de intervención en tiempo real. La visualización interactiva de los datos, tanto en mapas geográficos como en gráficos estadísticos, permite una comprensión clara y dinámica de los patrones de adopción digital a lo largo del territorio colombiano.

Sin embargo, el sistema presenta algunas limitaciones que deben tenerse en cuenta. La instalación y configuración inicial del Monitor de Inclusión Digital Colombia requieren ciertos conocimientos técnicos, particularmente en el uso de plataformas como Docker y AWS EC2. Esto podría ser un obstáculo para aquellos usuarios sin experiencia en estas tecnologías. Además, la efectividad del sistema depende de la calidad de los datos cargados. Si los datos no están correctamente preprocesados o contienen errores, el análisis generado podría no reflejar la realidad de la adopción digital en el país. Por último, el rendimiento del sistema también está condicionado por la infraestructura del servidor donde se ejecute, por lo que es necesario contar con una máquina con al menos 8GB de RAM para garantizar un funcionamiento adecuado.

Es fundamental que el usuario tenga en cuenta que para aprovechar todas las funcionalidades avanzadas del sistema, como los insights generados por la inteligencia artificial, será necesario configurar una clave de API de OpenAI. Además, debido a que el sistema está basado en tecnologías de contenedores, es crucial asegurarse de que el entorno de instalación cumpla con los requisitos técnicos establecidos (por ejemplo, la disponibilidad del puerto 5000 y una conexión a internet estable). Asimismo, dado que el sistema genera recomendaciones de políticas públicas basadas en los datos cargados, es importante que los usuarios verifiquen la calidad de los datos antes de tomar decisiones basadas

en los resultados generados. Un análisis erróneo debido a datos inexactos podría resultar en políticas incorrectas que no aborden adecuadamente la brecha digital.

2. Funcionamiento

Requisitos previos:

- Docker Desktop para Windows o Mac, o Docker Engine para sistemas Linux.
- 8GB de RAM mínimo recomendado.
- Puerto 5000 disponible para la conexión web.
- Clave de API de OpenAI (opcional para acceder a recomendaciones avanzadas basadas en inteligencia artificial).

Pasos para la instalación:

a) **Instalar Docker**: Descargue Docker Desktop desde el sitio web oficial Docker Desktop (para Windows o Mac) o instale Docker Engine (para Linux).

b) **Instalar en AWS EC2**:

b.1. Conéctese a su servidor EC2 utilizando el comando:
ssh -i tu-clave.pem ubuntu@tu-ip-ec2

b.2. Clonación del repositorio desde GitHub:
git clone <url-del-repositorio>
cd Apropiacion-Digital

b.3. Si se desea integrar la funcionalidad de IA, configure la clave de OpenAI de la siguiente forma:
export OPENAI_API_KEY="tu-clave-openai"

c) **Ejecutar el Script de Instalación**:

c.1. Asegúrese de que el script tenga permisos de ejecución:

chmod +x deploy_aws.sh
./deploy_aws.sh

d) **Acceder a la Aplicación**:

d.1. Una vez ejecutada la instalación, acceda a la aplicación desde el navegador:

http://tu-ip-ec2:5000

Personalización de Datos:

- Se permite la carga de nuevos conjuntos de datos a través de la interfaz web.
- Los archivos de datos se encuentran en la carpeta data/original/, donde es posible reemplazar los archivos existentes.
- Las variables de análisis pueden ajustarse modificando el archivo famd_clustering.py.

3. Casos de uso y paso a paso para el uso esperado

3.1 Cargar nuevos datos:

- En el menú principal, seleccione la opción Carga de Datos.
- Utilice el botón Subir archivo para seleccionar y cargar un nuevo conjunto de datos.
- Los clusters y los análisis se regenerarán automáticamente, actualizando los resultados en el dashboard.

3.2 Especificar una consulta:

- Diríjase a la sección Análisis de Segmentos para explorar los diferentes segmentos generados.
- Seleccione el segmento de interés y obtendrá un desglose detallado que incluirá características demográficas y tecnológicas asociadas a dicho segmento.

3.3 Elegir una forma de visualización:

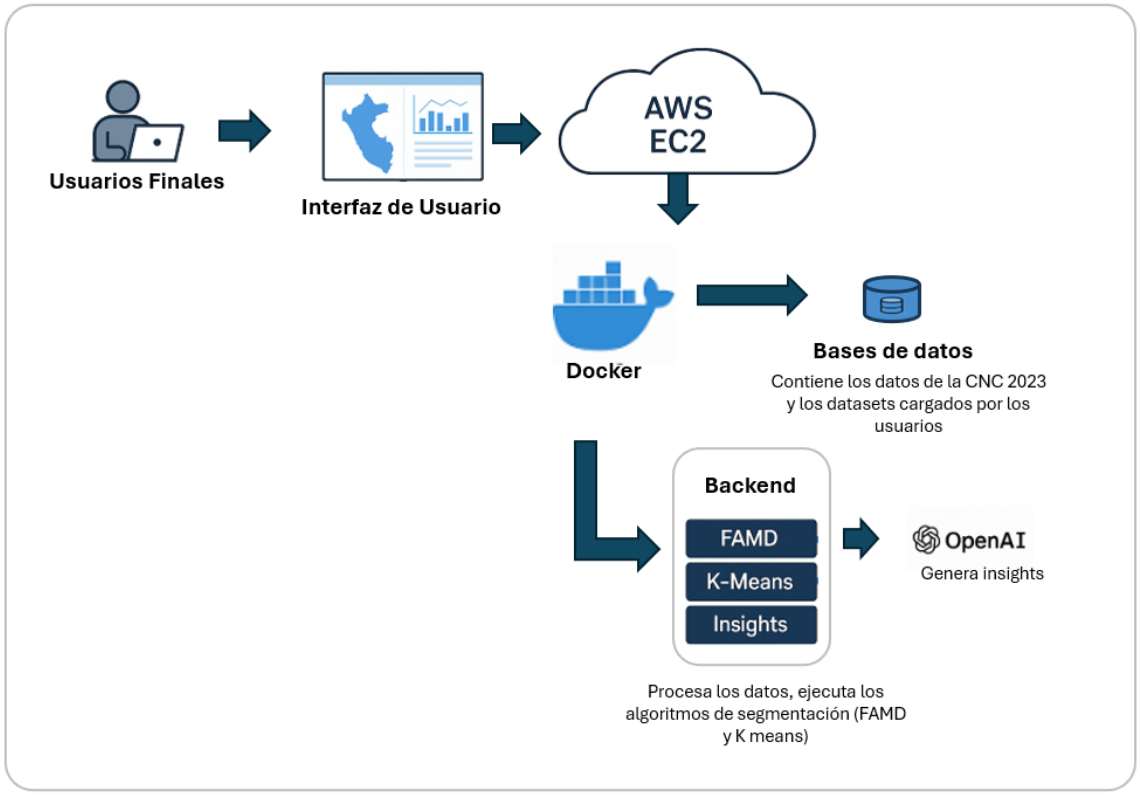
- En el Dashboard Interactivo, podrá seleccionar el tipo de visualización deseada: Mapas geográficos o Gráficos estadísticos.
- La visualización se actualizará en tiempo real conforme a las selecciones realizadas.
- Generar insights con IA: Acceda a la sección Insights de IA, donde se ofrecerá un análisis automático para cada segmento. La IA proporcionará recomendaciones orientadas a la formulación de políticas públicas y señalará oportunidades o desafíos específicos para cada segmento.

3.5 Visualización geográfica:

- En la sección Mapas Interactivos, se podrá explorar la distribución geográfica de la adopción digital en Colombia.
- Se podrá hacer clic en cada región del mapa para obtener detalles sobre la brecha digital en cada área específica.

Anexo técnico

1. Diagrama esquemático propuesto



2. Reporte técnico de experimento

El proceso de entrenamiento y calibración del modelo de segmentación se estructuró mediante la implementación y validación de tres enfoques experimentales, cada uno con arquitecturas distintas orientadas a identificar perfiles de apropiación digital en la población colombiana. Estos enfoques combinaron técnicas de reducción de dimensionalidad, agrupamiento no supervisado y evaluación interna mediante métricas específicas, preservando la coherencia metodológica en un entorno sin etiquetas de verdad.

Experimento 1 – LDA (Análisis Discriminante Lineal): Se evaluó un modelo supervisado de reducción de dimensionalidad, centrado en la separación lineal de grupos definidos previamente por niveles de apropiación digital. Este enfoque sirvió como punto de partida exploratorio para visualizar la capacidad discriminante de las variables más relevantes, aunque su utilidad se limitó a contextos donde se dispone de etiquetas.

Experimento 2 – K-Means + PCA / DBSCAN: El segundo modelo integró un pipeline de preprocesamiento, reducción de dimensionalidad mediante PCA, y agrupamiento con K-Means. La implementación permitió observar la varianza explicada acumulada y validar visualmente los grupos formados en el espacio proyectado. Posteriormente se aplicaron ajustes iterativos al número de componentes y valores de K, respaldados por la gráfica de inercia y el coeficiente de silueta, lo que permitió refinar los perfiles obtenidos. De forma complementaria, se utilizó DBSCAN como técnica de validación, con ajustes en los hiperparámetros eps y min_samples para identificar estructuras densas no captadas por K-Means.

Experimento 3 – MCA + GMM: Esta tercera arquitectura empleó el Análisis de Correspondencias Múltiples (MCA) para transformar variables categóricas, seguido de un modelo de mezclas gaussianas (GMM) para capturar pertenencias probabilísticas a distintos perfiles. Se seleccionaron componentes que explicaban más del 80 % de la inercia y se evaluaron distintas configuraciones de k (entre 2 y 10). La naturaleza probabilística del GMM resultó especialmente útil para identificar perfiles híbridos o con límites difusos, lo que enriqueció la comprensión de la apropiación digital en grupos intermedios.

Validación y métrica principal – Silhouette Score: Dado que el problema planteado es no supervisado, se prescindió de métricas tradicionales como RMSE o AUC. El Silhouette Score fue la métrica clave para evaluar la coherencia interna de los clústeres, midiendo la separación relativa entre grupos. Su uso fue transversal a todos los experimentos y resultó pertinente en contextos de alta variabilidad poblacional.

Proceso de ajuste y depuración de variables: Inicialmente se seleccionaron 12 variables representativas, derivadas de un proceso de reducción desde más de 800 columnas. Estas incluían indicadores sobre conectividad, tipo de dispositivos, percepción y frecuencia de uso de internet, y nivel educativo. Se aplicaron técnicas de correlación para evitar redundancias y se estandarizaron variables clave para corregir sesgos de escala.

Selección del modelo final: Aunque el modelo MCA + GMM ofreció valiosas interpretaciones en términos probabilísticos, se optó por el Modelo 2 (K-Means + PCA) como arquitectura principal de análisis. Esta decisión se sustentó en su mayor profundidad en la validación visual, la consistencia de los perfiles generados y la claridad en la representación de grupos diferenciados. El análisis de inercia y la maximización del Silhouette Score respaldaron la elección de K=5, valor que permitió lograr un balance adecuado entre granularidad y estabilidad de segmentación.

Perfiles identificados: Con base en la arquitectura seleccionada, se identificaron cinco segmentos poblacionales diferenciados en su nivel de apropiación digital:

- **Segmento 0 – Apropiación Crítica Alta:** Individuos con conectividad robusta, uso intensivo de internet, múltiples dispositivos y alto nivel educativo. Representan perfiles con liderazgo digital potencial.
- **Segmento 1 – Apropiación Funcional:** Usuarios con habilidades básicas y uso estable, aunque limitado a funciones específicas. Requieren estrategias de profundización de competencias.
- **Segmento 2 – Intermitencia Digital:** Acceso ocasional, dependencia de redes móviles y habilidades limitadas. Necesitan acompañamiento estructural y educativo.
- **Segmento 3 – Excluidos Conectados:** Tienen dispositivos, pero escasa apropiación digital o percepción negativa del internet. Requieren intervenciones pedagógicas y contextuales.
- **Segmento 4 – Apropiación Ausente:** Sin acceso, habilidades ni percepción de utilidad del internet. Constituyen el grupo más vulnerable, prioritario para políticas integrales de inclusión.

La implementación iterativa de modelos con distintos enfoques metodológicos permitió validar la solidez del proceso de segmentación. La arquitectura basada en PCA + K-Means demostró ser la más eficaz por su capacidad de representar perfiles diferenciados, su estabilidad ante múltiples configuraciones y su potencial de aplicación en escenarios reales de formulación de políticas públicas para cerrar la brecha digital. La elección del modelo final fue respaldada por análisis métricos y visuales consistentes, y su interpretación refuerza la utilidad del enfoque adoptado en contextos de alta heterogeneidad sociotecnológica.