Proyecto aplicado en analítica de datos

Reporte de implementación y experimentos

Integrantes:

Edwin Piñeros

José Contreras Arenas

Oscar Ardila

Wendy Galvis

Colombia, 7 de septiembre de 2025

1. Introducción

El objetivo de este proyecto es diseñar una herramienta dirigida a los analistas y tomadores de decisiones del Centro Nacional de Consultoría (CNC), que facilite la comprensión de los factores que limitan la inclusión digital, permita identificar a las poblaciones con mayor rezago y la exploración de escenarios de intervención respaldados por datos.

La necesidad central que busca atender el prototipo es transformar grandes volúmenes de información —proveniente de encuestas, indicadores oficiales y fuentes educativas— en conocimiento accionable para la formulación de políticas públicas. Para ello, se emplearán modelos de aprendizaje no supervisado que permitan identificar patrones en los datos y posteriormente visualizarlos en un tablero interactivo.

En este documento se presenta el reporte de implementación y experimentación de dichos modelos, incluyendo la verificación de supuestos, el tratamiento previo de los datos, los procesos de calibración, el análisis de resultados y el plan de implementación del prototipo.

2. Procesamiento y análisis exploratorio de los datos

Repositorio Github (ETL y análisis exploratorio de los datos): **https://github.com/WendyGalvisL/Adopci-n-digital.git**

Para este proyecto se trabajó principalmente con datos de la Encuesta de Apropiación Digital realizada por el CNC, complementados con información sobre acceso a internet proveniente de los resultados de las pruebas Saber 11 del ICFES y con datos de cobertura neta en educación del Ministerio de Educación Nacional. Estas dos últimas fuentes se incorporaron principalmente para enriquecer la caracterización de los grupos obtenidos en el análisis.

En cada fuente de información se realizó un proceso de exploración y limpieza, que incluyó la eliminación de columnas con altos porcentajes de valores faltantes, la imputación de datos ausentes y el ajuste de tipos de variables, entre otros. Adicionalmente, fue necesario consolidar las diferentes encuestas del CNC, lo que implicó unificar los nombres de las columnas antes de agregarlas en una única base.

Posteriormente, mediante el código DIVIPOLA del DANE como identificador único de cada municipio, se cruzó esta información con los datos del ICFES y del MEN.

Finalmente, sobre la base consolidada se realizó una revisión exhaustiva y una selección preliminar de las variables más relevantes para los modelos de clustering. El detalle completo del procesamiento de los datos se encuentra disponible en este enlace.

Luego de esto, se realizó el análisis exploratorio que permitió caracterizar de manera preliminar el conjunto de datos utilizado, compuesto por 7.662 registros y 41 variables de tipo numérico y categórico. Esta etapa resultó clave para comprender la estructura de la información, detectar inconsistencias y orientar las transformaciones posteriores requeridas para los modelos de clustering.

Calidad de los datos

Se identificaron 18 registros duplicados, los cuales representan un porcentaje reducido frente al total de observaciones. Asimismo, se evidenció un número importante de valores faltantes en algunas variables: por ejemplo, la pregunta P57 presenta un 92,98% de datos nulos, mientras que otras como P29 (36,44%) y P27 (21,98%) muestran porcentajes significativos de ausencia. En contraste, variables estructurales como MUNICIPIO_NOMBRE, Nombre Departamento, POBLACIÓN_ICFES o HOGARES_INTERNET no presentan datos faltantes, lo que asegura consistencia en los atributos centrales para el análisis

Estadísticos descriptivos

El conjunto de datos integra variables geográficas (GPSLAT, GPSLONG), sociodemográficas (EDAD, ESTRATO, NIVEL_PIRAMIDE), educativas (PROP_EDUC_5_16_MEN, POBLACIÓN_ICFES) y de infraestructura digital (HOGARES_INTERNET, TASA_INTERNET_ICFES). Entre los hallazgos relevantes se destacan:

- La edad promedio de los individuos es de 40 años, con una desviación estándar de 18, lo cual refleja una distribución amplia de la población encuestada.
- La variable estrato muestra una media de 2, con predominio de hogares en estratos bajos y medios.
- En términos de acceso digital, la tasa de internet ICFES registra un valor medio de 0,73 (73%), evidenciando una penetración heterogénea a nivel territorial.

Distribución y outliers

Los histogramas de las variables numéricas revelaron sesgos en indicadores como P7, P9, P27 y P64, asociados al uso de servicios y actividades digitales. La detección de outliers mediante el método del rango intercuartílico (IQR) indicó que variables como P7_1, P7, P64 y P29 concentran más del 15% de valores atípicos, lo que refleja la presencia de hogares con patrones de comportamiento marcadamente diferentes respecto al promedio nacional.

Variables categóricas y codificación

Preguntas de opción múltiple como P7, P15 y P27 presentaron alta cardinalidad y diversidad en las respuestas. Para estandarizar su uso en modelos de clustering, se implementaron variables dummies, conservando categorías críticas como:

- Servicios de telecomunicaciones (sí/no) a partir de P7.
- Disponibilidad de dispositivos en el hogar a partir de P15.
- Conexión del hogar a internet a partir de P27.

Este proceso permitió reducir la complejidad de la codificación original y asegurar la compatibilidad con métricas de distancia adecuadas, como la de Gower.

3. Requerimientos y selección de modelos

Los requerimientos de negocio de este proyecto se centran en identificar patrones de adopción digital en Colombia, relacionarlos con factores sociodemográficos y de infraestructura, apoyar la priorización de regiones con mayores brechas, y reconocer las variables más influyentes en la adopción digital. Estos requerimientos responden a preguntas críticas para el CNC, ya que permiten orientar la formulación de políticas y focalizar intervenciones en poblaciones y territorios específicos.

Para cumplir con estos objetivos, se propone el uso de modelos descriptivos basados en aprendizaje no supervisado, en particular técnicas de clustering. Estos modelos resultan adecuados porque no se cuenta con etiquetas previas que clasifiquen los niveles de adopción digital en los hogares; por tanto, el problema pertenece a la categoría de aprendizaje no supervisado. El clustering permite descubrir perfiles poblacionales y territoriales emergentes, ofreciendo una visión integrada de los determinantes sociales, económicos y tecnológicos de la adopción digital.

Aunque los modelos predictivos y prescriptivos pueden ser útiles en otros contextos, en este caso no responden de manera directa a la necesidad de comprender qué está ocurriendo actualmente y cómo se estructuran las brechas digitales. En este sentido,

los modelos descriptivos cumplen mejor la función de transformar datos en conocimiento accionable.

Finalmente, la integración de los resultados de clustering en un tablero interactivo permitirá que los analistas y tomadores de decisiones exploren los patrones identificados y utilicen la evidencia como insumo directo en la planificación estratégica.

4. Alternativas de modelos y técnicas

En este apartado se describen los modelos específicos que se probaron para el análisis y sus principales ventajas y desventajas. Como se mencionó anteriormente, todos son algoritmos de clustering, ya que permiten identificar patrones en la información sin necesidad de contar con etiquetas previas.

- **K-means:** Este algoritmo es uno de los métodos más utilizados en análisis de clustering debido a su simplicidad y eficiencia computacional. Su principal fortaleza radica en la capacidad de segmentar grandes volúmenes de datos en grupos definidos por la minimización de la varianza interna, lo que permite obtener clústers relativamente homogéneos y fáciles de interpretar. Entre sus ventajas se destaca la rapidez en la convergencia, así como la facilidad para ajustar parámetros y replicar resultados en diferentes experimentos. Sin embargo, presenta limitaciones relevantes: requiere definir previamente el número de clústers, lo que puede ser problemático en escenarios donde la estructura de los datos no es clara, y es sensible a la escala de las variables y a la presencia de outliers, que pueden distorsionar los centroides. Apesar de estas restricciones, K-means constituye una alternativa adecuada como punto de partida en este proyecto, ya que facilita la exploración inicial de patrones de adopción digital y permite generar una primera aproximación a la segmentación territorial y sociodemográfica.
- **K-medoides:** Este método de análisis de clustering es similar a K-means pero usando medoides (observaciones reales del conjunto de datos) como centros, lo que lo hace más robusto a outliers y muy adecuado para datos mixtos al permitir métricas como Gower y ponderación de variables. Por otro lado, el modelo es fácilmente interpretable, los clústeres se pueden describir con base en la observación real del medoide. Sin embargo, suele ser más costoso en tiempo y memoria (si se calcula una matriz de distancias), también exige fijar un numero de clúster (k) previamente y puede ser sensible a la métrica o pesos definidos. Esta alternativa puede ser relevante para el proyecto porque se ajusta de forma natural

a la estructura mixta de la base de datos (variables binarias, enteras y continuas) mediante el uso de disimilitudes y la ponderación de variables clave, es robusto a registros atípicos al utilizar medoides como centros y, por tanto, ofrece segmentos altamente interpretables.

- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise): es un algoritmo basado en densidades que permite identificar clusters de forma arbitraria, lo cual resulta especialmente útil cuando los datos presentan estructuras complejas. Una de sus principales ventajas es que puede detectar puntos que no pertenecen a ningún grupo, clasificándolos como ruido, lo que permite manejar de manera adecuada los datos atípicos. Además, no requiere que se especifique de antemano el número de clusters, lo que lo hace flexible frente a escenarios con estructuras desconocidas. Sin embargo, es importante tener en cuenta que DBSCAN es particularmente sensible a la elección de sus parámetros (eps y minPts), por lo que requiere una calibración cuidadosa para obtener resultados confiables. Asimismo, su aplicación puede implicar un alto costo computacional en datasets grandes, lo que debe considerarse al decidir su inclusión en la implementación final. Estas características hacen que DBSCAN sea una opción valiosa para identificar patrones complejos de adopción digital.
- **Clustering jerárquico:** Las jerarquías son estructuras que no sirven para organizar la informacion, mostrando relaciones anidadas y ayudan a comprender las relaciones entre sus elementos. Asi mismo, se utilizan para encontrar y formar clusteres. Entre las ventajas que podemos tener con este modelo, es que no nos limita a definir el número de grupos desde el principio, es posible realizar el proceso de definición de jerarquías de manera completa y luego decidir en que puntos realizar los cortes y determinar el número de clusteres 2,3 o más de acuerdo al análisis requerido. A través de la visualización de un dendrograma es posible la visualización de las relaciones jerárquicas, los grupos y subgrupos que se pueden presentar, con el cual se podrán identificar los patrones presentados. Aunque una de las desventajas es la escalabilidad limitada para conjuntos de datos grandes, su utilización es posible a través del uso de muestras o de reducción de dimensiones. Para nuestro caso es funcional si se quiere entender y presentar a los stakeholders un analisis exploratorio de cómo es la agrupación de municipios y zonas en términos de adopción digital.
- **Estadísticas descriptivas:** Luego de aplicar los modelos de clustering, es apropiado analizar también variables que no se utilizaron directamente para la creación de los clusters, pero que ayudan a describirlos. Estas estadísticas permiten caracterizar los grupos resultantes y generar visualizaciones útiles para el tablero,

facilitando la relación de cada cluster con características sociodemográficas u otras variables de interés. Esto permitiría complementar el análisis de los modelos y facilitaría la comprensión de las variables que influyen en la adopción digital y la priorización de regiones.

5. Verificación de supuestos y preparación de datos

Repositorio Github (incluye los códigos de cada implementación): https://github.com/WendyGalvisL/Adopci-n-digital.git

- **K-means:** Para aplicar este algoritmo fue necesario garantizar que los datos cumplieran con los supuestos básicos del modelo, en particular la homogeneidad en la escala de las variables y la ausencia de categorías sin tratamiento. En este sentido, se realizó una revisión exhaustiva de los valores faltantes, evidenciando variables con porcentajes elevados de nulos (por ejemplo, P57 con más del 90%), las cuales fueron descartadas para evitar sesgos en la segmentación. De igual forma, se identificaron y eliminaron registros duplicados. Dado que K-means se basa en la distancia euclidiana, fue imprescindible transformar preguntas de opción múltiple en indicadores binarios, creando variables *dummies* que permitieran estandarizar la representación de categorías como servicios de telecomunicaciones (P7) y equipamiento del hogar (P15). Las variables numéricas, como edad y proporción de hogares con internet, se conservaron en su escala original, pero al normalizarse dentro del cálculo de distancias evitaron que magnitudes mayores distorsionaran la asignación a clústers. Con estos ajustes, se aseguraron las condiciones para que K-means pudiera identificar agrupaciones coherentes, minimizando el riesgo de sesgos derivados de la heterogeneidad de los datos

- **K-medoides:** Para la implementación del modelo se realizó una verificación previa de supuestos y una depuración orientada al formato de los datos. En primer lugar, se restringió el conjunto de predictores a variables directamente vinculadas con conectividad y uso, indicadores binarios de *servicios/equipamiento* y la continua *dens_int*, excluyendo identificadores, textos y campos geográficos (lat/long) para evitar mezclar proximidad espacial con condiciones digitales y reducir la incidencia de valores faltantes. Las variables booleanas se codificaron explícitamente como 0/1 y se coercionaron los tipos a enteros; *dens_int* se convirtió a numérica y se imputó con la mediana ante ausencias. Dado que en esta implementación se utilizó K-medoids con distancia euclídea sobre una mezcla de binarias y una continua, se estandarizó *dens_int* (z-score) para evitar que su escala domine la disimilitud.

Adicionalmente, se habilitó ponderación de variables para reflejar prioridades analíticas (mayor peso a conectividad y uso). Se revisaron variables con mayor cantidad de registros nulos y se excluyeron de la modelación para no introducir ruido; del mismo modo, se descartaron campos redundantes o puramente administrativos. Finalmente, antes del ajuste se comprobó la consistencia básica de la base de datos (tipos, rangos plausibles y ausencia de categorías sin soporte), y tras el ajuste se aplica un umbral mínimo de tamaño por clúster coherente con la naturaleza robusta de atípicos del método al usar medoides (observaciones reales) como representantes.

- **DBSCAN:** Uno de los principales supuestos de este modelo es que las agrupaciones corresponden a regiones densas separadas por zonas de menor densidad; en consecuencia, su desempeño se ve limitado cuando los datos presentan niveles de densidad muy heterogéneos o cuando el ruido está ampliamente distribuido en toda la muestra. Para validar este supuesto en los datos utilizados, se calculó la matriz de distancias de Gower, adecuada para variables mixtas, y se aplicó el gráfico de distancias al vecino más cercano para identificar posibles umbrales de densidad. Posteriormente, se utilizó el método del codo para estimar el parámetro eps. Además de eps, la efectividad del modelo depende de la calibración de minPts, que es el número mínimo de puntos que deben estar dentro de la vecindad de un punto para que este sea considerado un punto central y, por tanto, se configure un clúster. Una configuración inapropiada puede producir demasiados grupos pequeños, un único clúster o clasificar excesivamente los datos como ruido. A diferencia de eps, no existe un método exacto para definir minPts; en este caso, dado que contamos con más de dos dimensiones en cada experimento, seguimos el criterio de Sander et al. (1998), quienes sugieren utilizar al menos 2*dimensiones.

  En cuanto al procesamiento previo de los datos, algunas variables de nuestro dataset presentan múltiples opciones de respuesta, lo cual constituye un primer aspecto a considerar, ya que el algoritmo se basa en distancias y este tipo de codificación puede afectar su desempeño. Para resolverlo, optamos por crear variables dummies y codificarlas en formato binario (1 y 0). Adicionalmente, los datos incluyen tanto variables categóricas como numéricas y dado que ciertas métricas de distancia, como la euclidiana, no resultan adecuadas para variables no numéricas, se decidió utilizar la distancia de Gower. Esta métrica permite trabajar con datos mixtos y, además, normaliza automáticamente las variables numéricas, evitando la necesidad de un escalamiento adicional. Con estos ajustes, aseguramos

que los datos cumplen los supuestos requeridos para aplicar DBSCAN de manera adecuada, permitiendo que la agrupación refleje patrones reales.

- **Clustering jerárquico:** En el modelo de clustering jerarquico es que las distancias son representativas, por lo tanto, se supone que si las observaciones están cerca una de otra deben ser similares, no se deben presentar variables dominantes o que puedan generar algún tipo de ruido. Este modelo al no manejar datos vacios, se deben realizar procesos de imputación de los datos (reemplazar o eliminar los datos faltantes). Para la selección de variables se deben conservar aquellas variables que nos permitan identificar niveles de adopción digital, y contengan información o datos demográficos, con esto evitamos la presencia de ruido o de datos que no nos den algún tipo de valor. Para evitar la existencia de variables dominantes se deben escalar o estandarizar los datos, teniendo en cuenta que el clustering se basa en distancias, y así evitamos que una variable tenga más peso que otra. Solo se incluirán variables numéricas, binarias para la realización del cálculo de las distancias, no se tendrán en cuentan aquellas variables categóricas (ejemplo nombre municipio y departamentos). Evaluaremos el número de clusteres a definir utilizando los métodos del codo y Silhouette.

6. Entrenamiento, calibración y métricas

Con el objetivo de identificar patrones de adopción digital en el país y las principales variables que influyen en esto, en este proyecto se optó por entrevistar cuatro tipos de modelos de clustering, probar con distintas variables y parámetros para identificar aquellas que nos arrojaran un mejor resultado. A continuación, se describen los experimentos explorados para cada tipo de modelo, así como su desempeño:

K-means (repositorio): Para el modelo K-means se desarrollaron tres fases de experimentación orientadas a garantizar la robustez de la segmentación y la coherencia con la pregunta de negocio. En el **Experimento 1** se planteó una línea base, evaluando configuraciones de k entre 4 y 12 con pesos iguales para todas las variables. La comparación de métricas internas (Silhouette, Calinski–Harabasz y Davies–Bouldin) mostró un desempeño óptimo en la configuración con **12 clústers** (Silhouette = 0,58), lo que indicó una buena separación entre grupos sin sacrificar cohesión interna. Los gráficos comparativos de Silhouette evidencian cómo este valor se incrementa consistentemente hasta alcanzar su máximo en k=12, justificando la elección inicial de granularidad.

Posteriormente, en el **Experimento 2** se incorporaron ponderaciones para resaltar el eje de adopción digital, asignando mayor peso a las variables de conexión, uso y densidad. El resultado produjo configuraciones más nítidas en la interpretación de perfiles (Profile Sharpness = 0,70), con una reducción a **8 clústers** que simplificó la lectura, aunque con un leve sacrificio en el índice de Silhouette (0,55). El gráfico de trade-off entre Silhouette y nitidez de perfil permitió visualizar claramente la tensión entre granularidad y simplicidad, mostrando a EXP1 como más detallado y a EXP2 como más explicativo.

Finalmente, el **Experimento 3** evaluó la estabilidad y coherencia territorial de los modelos candidatos. Se seleccionó la configuración de **12 clústers de EXP1**, la cual mantuvo un equilibrio adecuado entre separación y nitidez, alcanzando un **ARI promedio de 0,95** en 30 corridas y una entropía geográfica de 0,83. Esto validó que las agrupaciones no solo eran consistentes ante variaciones aleatorias, sino también territorialmente coherentes, lo que constituye un requisito clave para la focalización regional de políticas públicas.

En síntesis, el proceso de entrenamiento y calibración con K-means permitió explorar diferentes configuraciones y contrastar métricas de calidad y estabilidad. La evidencia gráfi ca mostró cómo la decisión final se apoyó en un balance entre granularidad analítica (EXP1, 12 clústers) y facilidad de interpretación (EXP2, 8 clústers), optando por la primera dado su mayor poder explicativo y coherencia territorial. Los resultados consolidan a K-means como un método adecuado para identificar **segmentos poblacionales diferenciados por acceso, equipamiento y confiabilidad digital**, respondiendo directamente al requerimiento de negocio de perfilar a las poblaciones rezagadas y priorizar territorios críticos para la intervención.

K-medoides (Repositorio): La preparación del modelo inicia con la selección de la variable centrada en adopción digital: indicadores binarios de servicios y uso (Servicios_Tele-comunicaciones_No/Si, Dispositivos_hogar_No, conexion_hogar_si, interrupciones_si, frec_uso_si) y la variable continua *dens_int*. Los booleanos se codificaron como 0/1 y *dens_int* se imputó por mediana y estandarizó; variables con nulos masivos quedaron excluidas.

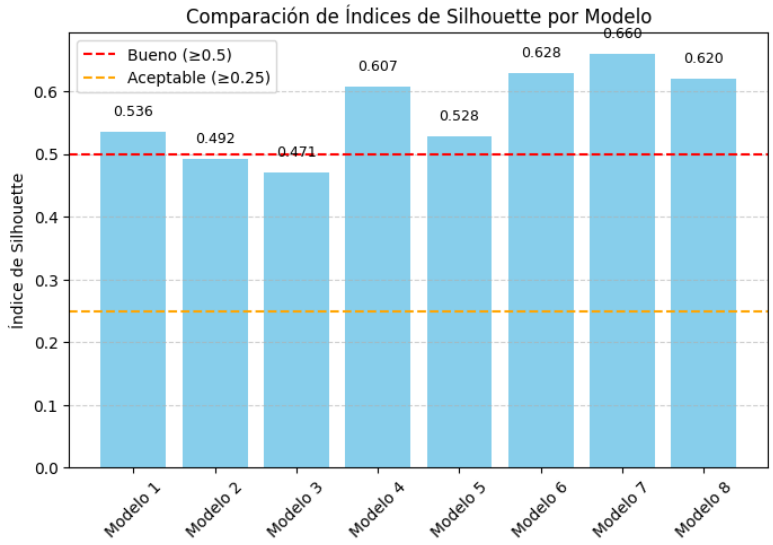
Para controlar sesgos, se definió un conjunto de entrenamiento y prueba estratificado por departamento: el modelo se ajustó en entrenamiento y, en prueba, cada observación se asignó al medoide más cercano para contrastar resultados fuera de muestra.

El entrenamiento, calibración y validación se ejecutaron en tres etapas complementarias. Inicialmente, se construyó una matriz de predictores con indicadores binarios de servicios y uso (0/1) y la variable continua dens_int estandarizada. Se empleó K-medoids con métrica euclidiana y una restricción de tamaño mínimo por clúster (≥ 2%) para evitar particiones triviales. Se exploraron valores de arreglos de cluster entre 4 y 12 con pesos uniformes y se compararon métricas internas: Silhouette, Calinski–Harabasz y Davies–Bouldin. La mejor configuración de esta fase fue k=10, con Silhouette ≈ 0,52 y distancia media de las variables binarias de 0,46, lo que sugiere clústeres diferenciados y cohesionados. Luego se realizó una ponderación orientada al negocio, para alinear el modelo con el eje de adopción digital, se duplicó el peso de conectividad y uso (Servicios_Telecomunicaciones_Si, conexion_hogar_si, frec_uso_si) y de dens_int. Esta calibración produjo perfiles más contrastados y separables, destacándose k=8 con Silhouette ≈ 0,55 y a la vez que mantuvo una distribución de tamaños equilibrada. Finalmente se realiza un experimento para verificar la robustez y coherencia territorial, se compararon los ganadores de EXP1 y EXP2 mediante una regla de decisión jerárquica, dando prioridad a Silhouette sobre ARI de estabilidad y nitidez. La solución ponderada con k=8 resultó preferible: mostró alta estabilidad entre corridas con submuestreo (ARI promedio elevado en múltiples semillas) y coherencia territorial evaluada con entropía geográfica normalizada (valores más bajos indican mayor concentración territorial del clúster). En conjunto, el proceso evidenció el clásico *trade-off* entre granularidad y legibilidad: la línea base ofrecía mayor detalle, mientras que la versión ponderada simplificó la interpretación sin perder separación. Dado su mejor equilibrio entre separación, estabilidad y claridad de perfiles, además de estar alineada con el objetivo sustantivo de medir adopción digital, K-medoids ponderado con k=8 se adoptó como configuración final para la segmentación socio-territorial.

DBSCAN: En este caso se realizaron ocho implementaciones del modelo con diferentes configuraciones de variables y parámetros. Dado que no contamos con etiquetas verdaderas, no es posible evaluar el desempeño con métricas supervisadas ni es necesario dividir los datos, por ello se utilizó el índice de Silhouette, ampliamente recomendado para modelos de clustering, estableciendo un umbral mínimo de 0.5 según los requerimientos iniciales. El detalle de todos los experimentos se puede ver aquí.

En la primera implementación se emplearon tres variables sociodemográficas (género, estrato y edad) y dos variables relacionadas con la adopción digital (frecuencia de uso de internet y servicios de telecomunicaciones utilizados). La selección de variables se basó en un análisis exploratorio previo y en el conocimiento del problema,

privilegiando aquellas que mostraban diferencias claras o que, por criterio, podían explicar la adopción digital. Para el cálculo de distancias se utilizó la métrica de Gower, dado que combina variables categóricas y numéricas y normaliza automáticamente las variables continuas. Los parámetros eps y minPts se definieron siguiendo los criterios metodológicos descritos en la sección anterior. El resultado de esta primera prueba arrojó un índice de Silhouette de 0.53, superando el umbral requerido.
Los experimentos siguientes siguieron un procedimiento similar, ajustando las variables y, en algunos casos, modificando los parámetros del modelo. Los resultados del desempeño se pueden observar en la siguiente gráfica:



En la gráfica se puede observar que el mejor desempeño se obtuvo con el modelo 7, que incluyó las variables área geográfica, edad, servicios de telecomunicaciones y cantidad de actividades realizadas en línea. Esta última se construyó a partir de una pregunta de opción múltiple en la encuesta, contabilizando el número de actividades como un proxy del nivel de adopción digital. El modelo generó 19 clústers, lo cual hace más complejo el perfilamiento de los usuarios. Finalmente, en el modelo 8 se aumentó el parámetro minPts con el objetivo de reducir la cantidad de agrupaciones. Este ajuste produjo 14 clústers con un índice de Silhouette de 0.62, lo que representa el mejor balance entre calidad de segmentación y viabilidad para el análisis.

Clustering jerárquico: Para la utilización y ejecución del modelo de clustering jerarquico se realizaron 5 experimentos en los cuales se realiza la seleccion de variables relevantes pensando en las preguntas de negocio. Durante el proceso se realiza la estandarización de las variables y evaluamos el modelo mediante el indice de Silhouette >= 0.5, asi mismo en el recorrido de cada experimento y con el mejor indice de Silhouette, definimos el mejor numero de clusteres (k) los cuales fueron probados

entre 2 y 12. Utilizaremos la distancia Gower ya que nuestro conjunto de datos cuenta con datos numericos y categoricos. La distancia Gower es una ventaja para el uso ya que se logra un mejor aporte de las variables teniendo en cuenta su naturaleza, evitando que esa variable quede diluida frente a las variables numéricas. A continuación, se explica cada uno de los experimentos los cuales se pueden encontrar en el siguiente link [Link Clustering Jerarquico]

Selección de variables para cada experimento:

Experimento	Variables	Enfoque
1	TASA_INTERNET_ICFES, dens_int, PERSONAS,AREA, Servicios_Telecomunicaciones_Si	Análisis de brecha digital, área
2	'PERSONAS', 'EDAD', 'REDAD', 'GENERO', 'ESTRATO', 'SERVICIOS_HOGAR', 'HOGARES_INTERNET'	Análisis Sociodemográfico, estructura de los hogares y acceso a internet
3	'PERSONAS', 'ESTRATO', 'SERVICIOS_HOGAR', 'HOGARES_INTERNET','FRECUENCIA_INTERNET', 'NIVEL_USUARIO_USO_INTERNET'	Análisis Sociodemográfico, estructura de los hogares, uso del internet y habilidades en el uso del internet
4	'TASA_INTERNET_ICFES', 'dens_int', 'Servicios_Telecomunicaciones_Si', 'conexion_hogar_si', 'interrupciones_si'	Análisis de Infraestructura, Acceso y calidad del servicio
5	"TASA_INTERNET_ICFES","dens_int","PERSONAS","AREA","Servicios_Telecomunicaciones_Si", 'conexion_hogar_si','NIVEL_USUARIO_USO_INTERNET'	Combinacion Analisis de acceso, infraestructura y habilidades digitales

Resultado de la ejecución de cada experimento

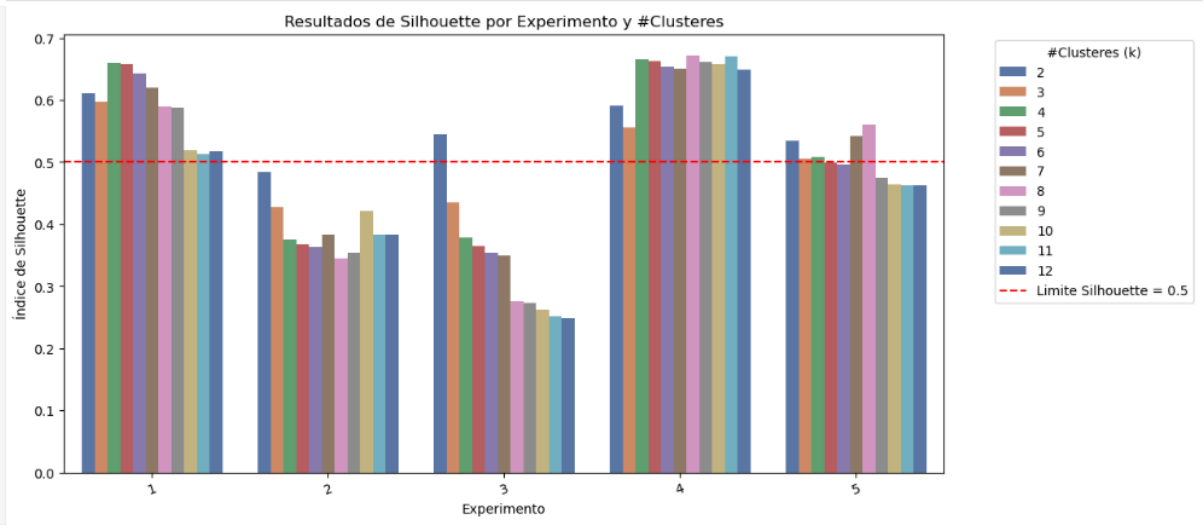


Tabla de resultados Indice de Silhouette

Experimento	# Clusters(k)	Silhouette
4	8	0.672677
1	4	0.660161
5	8	0.559646
3	2	0.545214
2	2	0.484212

Los experimentos 1 y 4 teniendo en cuenta que los resultados de los indices de Silhouette están por encima del límite definido de 0.5 son los que generan segmentaciones más sólidas. Los experimentos 2 y 3 no son tan confiables porque sus resultados están por debajo del límite de 0.5 y el experimento 5 son pocos los valores de k que se acercan al límite del índice de Silhouette.

7. Resultados y análisis

En este apartado, se realiza un análisis de las mejores alternativas por cada modelo y se define cuales alternativas funcionan mejor para el problema de negocio:

- **K-means:** El modelo K-means permitió obtener una segmentación robusta de la población en 12 clústers, resultado de la configuración seleccionada en los experimentos y validada por métricas internas (Silhouette = 0,58; Calinski–Harabasz > 5700; Davies–Bouldin < 0,80), estabilidad (ARI promedio = 0,95) y coherencia territorial (entropía geográfica = 0,83). Los centroides obtenidos reflejaron tres grandes tipologías: clústers con brechas de acceso y equipamiento (C3, C4, C9, C10), clústers con problemas de calidad y continuidad del servicio (C1, C2, C6, C8), y clústers de conectados estables o en transición (C0, C5, C7, C11). La representación geográfica mostró patrones regionales claros: Bogotá y Antioquia concentrados en clústers conectados con problemas de continuidad, la Costa Caribe en clústers con rezagos de acceso y dispositivos, y regiones periféricas en clústers donde persisten dificultades estructurales de cobertura.

  La principal fortaleza de K-means radica en su alta interpretabilidad, lo que lo convierte en un modelo idóneo para la toma de decisiones. Cada clúster se describe mediante variables observables como acceso, dispositivos, interrupciones y densidad digital, facilitando la comunicación de resultados y la definición de estrategias focalizadas. A diferencia de otros métodos, K-means ofrece segmentos estables y territorialmente coherentes que permiten derivar acciones concretas: subsidios a dispositivos, inversión en infraestructura, programas de alfabetización digital o migración hacia conexiones fijas. En este sentido, los resultados

consolidan a K-means como la técnica más adecuada para identificar poblaciones rezagadas, priorizar territorios críticos y orientar intervenciones de inclusión digital.

- **K-medoides:** Con base en los experimentos, K-medoids ponderado con k=8 es una alternativa que se ajusta a los requerimientos de segmentación interpretable y operativa del proyecto. Se obtuvieron las siguientes métricas de desempeño: Silhouette de 0,55, Calinski–Harabasz de 6.203, Davies–Bouldin de 0,725 y tamaños de clúster entre ≈5% y 23%. El modelo ponderado superó el no ponderado (mejor en k=10 y Silhouette ≈0,523) al ofrecer perfiles más nítidos y menos solapados; la ponderación sobre Servicios_Telecomunicaciones_Si, conexion_hogar_si, frec_uso_si y dens_int fortaleció el eje de adopción digital sin sacrificar estabilidad. En cuanto a los perfiles, el modelo identifica cinco tipologías operativas: (i) Sin servicios de telecomunicaciones (clúster 5; 4,6%), grupo crítico con ausencia total de servicio y alta carencia de dispositivos; (ii) Sin conexión en el hogar (clústeres 2 y 3; 13,1% y 7,6%), con escasez de dispositivos y contraste de densidad (baja en C2, alta en C3), lo que sugiere barreras de adopción más que de oferta en este último; (iii) Conectados con mala calidad (clústeres 0 y 7; 23,4% y 18,1%), con interrupciones universales pese a contar con conexión y equipamiento; (iv) Conectados estables/satisfechos (clústeres 1 y 4; 9,8% y 17,9%), con conexión sin interrupciones y diferencias de densidad (media en C1, baja en C4); y (v) Conectados intermitentes (clúster 6; 5,4%), con interrupciones en torno al 57% en zonas de mayor densidad, compatible con congestión de red. La calidad de separación intra-clúster respalda estas tipologías (mejor definición en C0, C4, C2, C5 y C7; menor nitidez relativa en C1, C3 y C6), lo que permite orientar intervenciones diferenciadas enfocadas en acceso básico, dotación, mejora de calidad y gestión de capacidad según el perfil identificado.

- **DBSCAN:** tal como se mencionó en el punto anterior, para esta implementación, se escogió el modelo que generó un mejor balance entre la cantidad de agrupaciones y el desempeño. Particularmente, el modelo escogido tiene 14 clústers con un índice de Silhouette de 0.62. Ahora, para lograr caracterizar los perfiles, se creó una tabla que incluyó los valores promedio o las modas asociadas a las variables más relevantes, aquí se incluyeron tanto las variables de segmentación como otras variables de contexto, asimismo, se crearon algunos indicadores que podían ayudar a comprender mejor el resultado de la tabla (ej: tasa de acceso a internet, tasa de cobertura neta en educación, porcentaje de

mujeres en el cluster, entre otras). La tabla con los resultados se puede ver en este link, así como la descripción detallada de cada uno de los pérfiles identificados.
Como resultado de lo anterior, se puede concluir que el modelo permitió identificar algunos grupos con sentido, diferenciando a los desconectados, a los parcialmente conectados y a los usuarios con mayor diversidad de servicios y actividades en línea. Sin embargo, la segmentación resultante se centró principalmente en variables sociodemográficas como edad, área de residencia y género, lo que indica que el algoritmo priorizó esas diferencias más que las relacionadas con los patrones de uso de internet y la apropiación digital, que eran el foco principal del trabajo.

- **Clustering jerárquico:** El mejor resultado fue dado en el experimento número 4, el cual tuvo un índice de **Silhouette 0.672677 y** un valor de k=8 el número de clusteres. Las variables relacionadas en el modelo son las siguientes **'TASA_INTERNET_ICFES', 'dens_int', 'Servicios_Telecomunicaciones_Si', 'conexion_hogar_si', 'interrupciones_si,** con las cuales podemos realizar un Análisis de Infraestructura, acceso y calidad del servicio. Procesando con estes parámetros el modelo de clustering los patrones y perfiles econtrados son los siguientes:

Perfiles de clústeres

Cluster	Tamaño	TASA_INTERNET_ICFES	dens_int	PERSONAS	AREA	Servicios_Telecomunicaciones_Si	conexion_hogar_si \	NIVEL_USUARIO_USO_INTERNET
1	2239	0.73	0.73	3.64	1.14	1.00	1.0	1.44
2	76	0.49	0.49	3.95	1.29	0.41	1.0	0.50
3	3351	0.77	0.77	3.61	1.04	1.00	1.0	1.50
4	30	0.76	0.76	3.70	1.07	1.00	0.0	0.97
5	66	0.55	0.55	4.32	1.20	0.29	1.0	0.50
6	7	0.59	0.59	3.86	1.00	0.14	0.0	0.43
7	1401	0.72	0.72	3.15	1.18	1.00	0.0	0.45
8	492	0.58	0.58	3.26	1.27	0.28	0.0	0.25

- **Cluster 1 (2.239 casos)**
  - Alta tasa de internet (0.73), todos con conexión en el hogar y telecomunicaciones.
  - Nivel de usuario alto (1.44), sin interrupciones.

    - Usuarios conectados y estables.
- **Cluster 2 (76 casos)**
    - Tasa media de internet (0.49), 41% con servicios, todos conectados al hogar.
    - Nivel bajo de usuario (0.50).
    - Conectados, pero con infraestructura limitada.
- **Cluster 3 (3.351 casos, el mayor)**
    - Alta tasa de internet (0.77), 100% conectados y con telecomunicaciones.
    - Nivel de usuario muy alto (1.50).
    - Usuarios avanzados, consolidados.
- **Cluster 4 (30 casos)**
    - Alta tasa de internet, 100% con servicios, pero sin conexión en el hogar.
    - Nivel de usuario intermedio (0.97).
    - Acceso parcial fuera del hogar (ej. redes públicas, cafés internet).
- **Clúster 5 (66 casos)**
    - Tasa media (0.55), baja disponibilidad de servicios (29%), con conexión en el hogar.
    - Nivel bajo de usuario (0.50).
    - Conectados en condiciones precarias.
- **Clúster 6 (7 casos, muy pequeño)**
    - Muy baja disponibilidad de servicios (14%), sin conexión en el hogar.
    - Nivel muy bajo de usuario (0.43).
    - Exclusión digital crítica (minoría, pero políticamente relevante).
- **Clúster 7 (1.401 casos)**
    - Alta tasa de internet (0.72), todos con servicios, pero sin conexión en el hogar.
    - Nivel bajo de usuario (0.45).
    - Usuarios dependientes de acceso externo/público.
- **Clúster 8 (492 casos)**
    - Tasa media (0.58), baja disponibilidad de servicios (28%), sin conexión en el hogar.
    - Nivel bajo de usuario (0.25).
    - Usuarios desconectados estructuralmente.

Conclusión

- Clústeres 1 y 3 Mayoría Conectados, mejora en la conexión de internet

- Clúster 6,7 y 8 falta de conexión en el hogar y desconexión estructural mejora en infraestructura zonas más vulnerables.

En términos generales, los resultados obtenidos muestran que los modelos de clustering permiten cumplir de manera satisfactoria los requerimientos del proyecto. El modelo K-means se destaca como la alternativa principal, ya que logra un balance óptimo entre desempeño técnico e interpretabilidad, ofreciendo clústeres estables, territorialmente coherentes y directamente vinculados con factores sociodemográficos y de infraestructura. Esto lo convierte en una herramienta robusta para identificar perfiles diferenciados de adopción digital, priorizar territorios críticos y orientar decisiones estratégicas de política pública, ajustándose de manera integral a los objetivos de negocio y a las métricas de evaluación definidas.

8. Plan de implementación del prototipo

Con el fin de completar la construcción del prototipo funcional, se ha definido un plan de implementación dividido en cuatro fases. Cada fase se describirá a continuación y responde a componentes y requerimientos aún pendientes por desarrollar.

Fase 1 — Empaquetado del pipeline (ETL + modelo) [hasta 12 sep].

ETL ejecutable con el mismo preprocesamiento usado en los experimentos (selección de features, dummies/normalización donde aplique, escalado de dens_int) y chequeos de calidad.

Entrenamiento/scoring ejecutables que llaman al bloque de malla/selección ya implementada (run_modelo_grid) y exportan labels, centroides, tamaños y métricas a /results.

Entregables: etl_build.py, train_modelo.py, score_modelo.py, results/{centroides,sizes,metrics,labels}.csv.

Fase 2 — Visualizaciones y tablero mínimo [13–17 sep].

Generación automática de figuras Silhouette vs k, CH, DB y trade-off Silhouette–Nitidez, más perfiles: edad (boxplots), estrato (barras apiladas), nivel de pirámide (heatmap) y dens_int; además, mapa por clúster dominante por departamento alimentado con los labels exportados.

Entregables: viz_report.py, carpeta assets/fig_*, y notebook/HTML con las figuras embebidas para el informe.

Fase 3 — Documentación y tabla diligenciada [18–20 sep].

Manual de uso.

Tabla diligenciada con centroides, tamaños, top territorios y métricas.

Entregables: docs/manual_usuario.pdf, docs/tabla_diligenciada.xlsx.

Fase 4 — QA y entrega ejecutiva [21–23 sep].

Pruebas de reproducibilidad, checklist de completitud y empaque final (ejecutables, código, manual, tabla).

Referencias

Sander, J., Ester, M., Kriegel, HP. et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery 2, 169–194 (1998). https://doi.org/10.1023/A:1009745219419