Proyecto:

MODELO DE AGRUPACIÓN DEL DESARROLLO SOCIOECONÓMICO DE LAS FAMILIAS EN COLOMBIA Y SU RELACIÓN CON LA CONECTIVIDAD A INTERNET BASADO EN IA.

Tripulantes - Grupo No. 1

María Andréa Patiño Ruiz

Maria Paula Rodriguez Castrillón

Jhon James Giraldo Patiño

Juan Pablo Granada Castaño

José Mauricio Arenas Cárdenas

Ejecutores:

- **Natalia Betancur Herrera.**
- **Frank Yesid Zapata Castaño.**
- **Margarita María Orozco.**

Manizales, octubre 2024

TABLA DE CONTENIDO

MODELO DE AGRUPACIÓN DEL DESARROLLO SOCIOECONÓMICO DE LAS FAMILIAS EN COLOMBIA Y SU RELACIÓN CON LA CONECTIVIDAD A INTERNET BASADO EN IA.

I. Introducción.

En la actualidad, la conectividad a internet es una de las herramientas fundamentales que influye en el desarrollo de las personas y las comunidades. En Colombia, la conexión a internet no solo permite el acceso a información y conocimientos, sino que se convierte en un facilitador para mejorar la calidad de vida de las familias, influir en su desarrollo socioeconómico y abrir oportunidades en diversas áreas, como la educación, el empleo y la productividad. Sin embargo, el "77,2% de los hogares colombianos tenían acceso a internet en 2022; 88,1% de la población en cabeceras municipales tenía acceso, frente al 44,4% en centros poblados y rural disperso, lo que evidencia una brecha digital significativa entre zonas urbanas y rurales". (DNP: 2023)

Muchas regiones del país enfrentan limitaciones en acceso a internet, lo que genera disparidades en el aprovechamiento de estas oportunidades. Esta brecha digital se relaciona con factores como el nivel de educación, los ingresos familiares, la tasa de analfabetismo y las oportunidades de empleo, los cuales en conjunto impactan en la calidad de vida y el desarrollo en los territorios.

La importancia del internet es especialmente evidente en el ámbito educativo como herramienta que facilita los procesos de aprendizaje, toda vez que según cifras del DNP "el 72,1% de los estudiantes usaron internet para actividades educativas en 2021" y la facilidad en el "acceso a internet puede mejorar el rendimiento académico en un 20%" (DNP:2023). Adicionalmente, permite a los estudiantes y docentes acceder a plataformas como YouTube, bibliotecas digitales y otros recursos que enriquecen el aprendizaje.

En términos de productividad, el acceso a nuevas tecnologías y sistemas de información permiten mejoras sustanciales en sectores como el agrícola, donde la conexión permite implementar sistemas de producción más eficientes y sostenibles. De igual forma, nuevas alternativas laborales como el teletrabajo e incluso las posibilidades de acceso a fuentes de empleo directa a través de plataformas de búsqueda de empleo.

Esta investigación se centrará en un enfoque analítico descriptivo del fenómeno social, empleando modelos no supervisados como K Means y jerárquicos, adecuados a la naturaleza de los datos disponibles. A partir de los conocimientos adquiridos en el “Bootcamp de Inteligencia Artificial” nivel exploradores impartido por el MINTIC, el proyecto estará centrado en un modelo de agrupación con respecto al acceso a internet en las comunidades del país y su impacto en el desarrollo socioeconómico de las poblaciones en Colombia, utilizando como base los datos de la “Encuesta de Calidad de Vida” realizada por el DANE en el año 2023. Para enriquecer el análisis se revisaron fuentes nacionales e internacionales, a partir de fuentes primarias sobre la calidad de vida de las poblaciones objeto de esta investigación.

II. Planteamiento del Problema

Las personas que no tienen acceso a internet pueden enfrentar limitaciones que afectan su calidad de vida y restringen oportunidades clave para el desarrollo socioeconómico. Esta falta de acceso influye directamente en una serie de variables interrelacionadas, tales como el nivel de estudios, el analfabetismo, los ingresos del hogar y las oportunidades de empleabilidad.

Este proyecto tiene como propósito construir un modelo de agrupación de familias en Colombia, basado en patrones que integren las variables de conectividad a internet y desarrollo socioeconómico, con el fin de guiar la priorización de políticas de conectividad enfocadas en mejorar las condiciones de vida en las diversas regiones del país. A través del uso de inteligencia artificial, se busca identificar los criterios de mayor impacto en estas condiciones, permitiendo una intervención estratégica y priorizada que cierre la brecha digital y fomente el crecimiento inclusivo en todo el país.

III. OBJETIVOS.

- **Objetivo general**.

Diseñar un modelo de agrupación de familias en Colombia, basado en patrones que integren las variables de conectividad a internet y desarrollo socioeconómico.

- **Objetivos específicos.**

➢ Analizar la información de la “Encuesta de Calidad de Vida” realizada por el DANE en 2023, identificando y seleccionando las variables relevantes de conectividad a internet y desarrollo socioeconómico que impactan la calidad de vida en diferentes regiones de Colombia.

➢ Realizar un análisis descriptivo de las variables, destacando los patrones socioeconómicos y de conectividad que permiten caracterizar a las familias en función de su nivel de acceso a internet y su situación socioeconómica.

➢ Seleccionar y aplicar modelos de agrupación no supervisados, como K Means y algoritmos jerárquicos, evaluando la adecuación de cada uno para identificar patrones de desarrollo y conectividad en las comunidades del país.

➢ Implementar un modelo de agrupación efectivo que permita clasificar a las familias en distintos grupos de desarrollo y conectividad, facilitando una interpretación práctica de los resultados.

➢ Proponer recomendaciones estratégicas basadas en los resultados del modelo, orientadas a la priorización de políticas de conectividad en las

regiones con mayores necesidades, con el objetivo de cerrar la brecha digital y promover el desarrollo inclusivo en todo el país.

IV. JUSTIFICACIÓN.

Este proyecto se justifica en la necesidad de reducir las desigualdades en acceso a internet en Colombia, donde “solo 60,5% de población colombiana tiene acceso a este servicio.” El país está lejos de llegar al nivel de países desarrollados como “Corea del Sur quien lidera el ranking con un nivel de cobertura de 99,9%. A este le sigue muy de cerca Noruega con 99%”.(La República: 2023)

Los hogares carecen de conexión, limitando su acceso a recursos educativos, laborales y productivos. Estas desigualdades son más evidentes en las zonas rurales, donde la conectividad es inferior al 25%, en contraste con un 65% en áreas urbanas (DNP: 2023). La brecha digital tiene implicaciones directas en el desarrollo socioeconómico, dado que el acceso a internet es una herramienta clave que impacta positivamente en la educación y en la productividad laboral de las comunidades, contribuyendo al desarrollo inclusivo y sostenible del país.

Este proyecto, busca desarrollar un modelo de inteligencia artificial orientado a la agrupación de datos, permitiendo identificar patrones clave, facilitando la definición de criterios de prioridad para apoyar políticas y estrategias de acceso digital en sectores vulnerables. *A su vez, busca ofrecer un modelo de agrupación que respalde el diseño de políticas inclusivas, promoviendo así una Colombia más conectada y equitativa*, con impacto directo en la mejora de indicadores de educación, empleo y productividad.

V. MARCO CONTEXTUAL

El acceso a internet es hoy un derecho que guarda relación directa con la calidad de vida de las personas, dado que habilita el acceso a servicios esenciales como la educación, la salud y el empleo. Esto se alinea con el artículo 25 de la Declaración Universal de Derechos Humanos de la ONU, que establece que toda persona tiene derecho a “un nivel de vida adecuado que le asegure, así como a su familia, la salud y el bienestar” (Naciones Unidas, 1948). La conexión a internet es, por tanto, un componente clave para que las personas puedan alcanzar este nivel de vida adecuado en la era digital, ya que facilita la igualdad de oportunidades y el ejercicio de derechos fundamentales.

De la misma manera, la UNESCO ha destacado que la inclusión digital es esencial para el desarrollo humano, afirmando que la falta de acceso a internet exacerba las desigualdades sociales y económicas. Según esta organización, el acceso a internet es una vía para mejorar la calidad de vida, pues permite a las personas aprovechar recursos educativos y laborales, factores críticos en la reducción de la pobreza y la promoción del desarrollo equitativo (UNESCO, 2021). La digitalización, por tanto, no sólo transforma las sociedades, sino que también abre la puerta a la justicia social y al progreso inclusivo.

La OCDE también resalta la relevancia de la conectividad para mejorar las condiciones de vida, especialmente |en países en desarrollo. En su informe Digital Economy Outlook, subraya que la falta de acceso adecuado a internet genera barreras al crecimiento económico y al bienestar general, destacando que la brecha digital es un obstáculo significativo para que las poblaciones vulnerables puedan alcanzar el desarrollo sostenible (OCDE, 2022). La inclusión digital, promovida a través de políticas y programas específicos, es vista como una estrategia fundamental para cerrar estas brechas y fortalecer el desarrollo socioeconómico en diversas regiones del mundo.

- **Contexto Histórico y Social**

La brecha digital es un problema persistente en muchas regiones del mundo, y en América Latina, Colombia es uno de los países donde esta desigualdad es notable. A lo largo de los años, el acceso a internet se ha convertido en un servicio fundamental que impacta directamente la calidad de vida de las personas,

influenciando el acceso a la educación, el empleo, la salud y otros derechos básicos.

Sin embargo, factores históricos de desigualdad social y económica han contribuido a que el acceso a las tecnologías sea desigual, principalmente en zonas rurales y en comunidades de bajos ingresos.

- **Situación Actual de la Conectividad en Colombia**

Colombia ha avanzado significativamente en la expansión de la infraestructura de internet, especialmente en áreas urbanas. Según datos recientes del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC: 2023), en los últimos años ha habido un crecimiento en la cobertura y el acceso a internet. Sin embargo, todavía existen barreras para la conectividad en áreas rurales, donde los costos, la falta de infraestructura y el limitado acceso a dispositivos adecuados son problemas importantes. Estas barreras han ampliado la brecha digital, que limita las oportunidades de desarrollo económico y social en las regiones menos conectadas.

- **Impacto Socioeconómico de la Conectividad a Internet**

El acceso a internet influye en el desarrollo socioeconómico de las familias, ya que facilita el acceso a servicios educativos, oportunidades laborales y recursos de salud. En este sentido, el acceso a internet puede convertirse en un catalizador del cambio, promoviendo el desarrollo humano y la equidad. Estudios recientes muestran que las familias con acceso a internet suelen tener mayores oportunidades de empleo, mejor rendimiento académico y mejores condiciones de salud, lo que impacta positivamente en sus ingresos y calidad de vida (UNESCO: 2021). Por lo tanto, entender la correlación entre el nivel socioeconómico y el acceso a internet permite identificar tanto los grupos vulnerables como las áreas de oportunidad para políticas públicas y programas de apoyo.

- **Esfuerzos y Políticas Públicas**

En Colombia, el gobierno ha implementado varios programas para fomentar la conectividad, especialmente en áreas rurales. El programa "Conectividad Total" y las iniciativas de inclusión digital son ejemplos de los esfuerzos por llevar internet a comunidades marginadas. Además, las políticas de subsidios a la conexión en hogares de bajos ingresos son intentos de reducir la brecha digital. Sin embargo,

las dificultades de implementación, junto con la falta de recursos en ciertas zonas, han limitado el alcance de estos programas, lo que genera una necesidad de estrategias basadas en datos para dirigir mejor los esfuerzos hacia los grupos más vulnerables.

- **Importancia del Proyecto en el Contexto Actual**

Este proyecto se enmarca en un contexto donde la transformación digital es fundamental para la inclusión social y el crecimiento económico. Entender la relación entre el nivel socioeconómico y la conectividad a internet permite identificar patrones y grupos que necesitan apoyo específico. La segmentación mediante el uso de técnicas de inteligencia artificial, como el algoritmo K-means, proporciona una herramienta valiosa para generar conocimiento accionable sobre los factores que inciden en la brecha digital. Estos datos pueden ser utilizados por los responsables de políticas y las organizaciones para desarrollar estrategias focalizadas que promuevan un desarrollo socioeconómico más equitativo.

VI. ALCANCE

El proyecto se enmarca dentro del contexto del bootcamp de Talento Tech del Ministerio de las TIC, que tiene como objetivo desarrollar habilidades en el ámbito tecnológico y promover la inclusión digital. A continuación, se detallan los límites y el enfoque del proyecto:

1. Duración del Proyecto

- El proyecto se desarrollará a partir del 16 de septiembre hasta el 9 de noviembre, en el marco del bootcamp. Esto implica que todas las actividades de diseño y análisis deberán ser completadas en este periodo.

2. Limitaciones

- **No Implementación:** El modelo creado no se implementará ni se desplegará en entornos reales, por lo que no se realizarán pruebas o validaciones en campo.
- **Foco Teórico:** El enfoque del proyecto será principalmente teórico y analítico, centrado en la construcción del modelo y la interpretación de los resultados, sin intervención directa en las comunidades.
- **Recursos y Herramientas:** El uso de herramientas de inteligencia artificial se limitará a la simulación y análisis de datos ya disponibles, sin necesidad de integrar sistemas en tiempo real.

3. Resultados Esperados

- **Informe de Resultados:** Al final del bootcamp, se presentará un informe que resuma los hallazgos del análisis, el diseño del modelo de agrupación y recomendaciones para futuras investigaciones o políticas.
- **Presentación:** Se organizará una presentación del modelo y los resultados ante los participantes del bootcamp y otros interesados, destacando la importancia de la conectividad para el desarrollo socioeconómico.

4. Beneficios Potenciales

- **Conocimiento Adquirido:** Los participantes del bootcamp ganarán experiencia práctica en el análisis de datos y el diseño de modelos, lo que fortalecerá sus competencias técnicas.
- **Base para Futuras Políticas:** Aunque no se implementará el modelo, los hallazgos pueden servir como base teórica para la formulación de políticas de conectividad en Colombia.

Este alcance busca abordar el problema del acceso a internet desde una perspectiva analítica y formativa, sin intervención directa, en línea con los objetivos del bootcamp de Talento Tech.

VII. Aspectos Técnicos y Metodológicos.

Etapas del proyecto.

El proyecto se desarrollará a través de las siguientes actividades y etapas clave, con el objetivo de comprender y analizar el impacto de la conectividad a internet en el desarrollo socioeconómico de las familias en Colombia, mediante un modelo de agrupación no supervisado enfocado en las distintas regiones del país:

- **Recolección y procesamiento de datos.**
  En primera instancia se realizará la consulta a fuentes primarias de información como repositorios especializados, centros de estudio y estadística, tales como la Encuesta de Calidad de Vida del DANE -ECV- y otros estudios económicos y sociales, los cuales presentan información de variables relacionadas con el acceso a internet, niveles de educación, condiciones de empleabilidad y calidad de vida de las familias en las diversas regiones de Colombia. Este proceso incluye el aprestamiento, limpieza, transformación y normalización de los datos para asegurar su compatibilidad y precisión en el análisis.
- **Análisis exploratorio de datos**
  Esta fase corresponde al análisis exploratorio de los resultados de la -ECV 2023- para identificar patrones, relaciones significativas, varianzas, dispersión de los datos, especialmente entre el acceso a internet y el desarrollo socioeconómico.

  Con base en métricas estadísticas se determinará el comportamiento de las variables, tales como la penetración de internet, ingresos familiares, tasas de desempleo y pobreza percibida, y su variación entre las regiones y cómo estas interactúan entre sí.

- **Desarrollo del modelo de agrupación no supervisado**
  Utilizando técnicas de machine learning, se implementará un modelo de agrupación no supervisado, como el algoritmo K-Means y métodos jerárquicos, para identificar grupos de regiones con características socioeconómicas y niveles de conectividad similares. Este modelo permitirá clasificar las regiones de acuerdo con patrones de conectividad y

desarrollo, estableciendo la correlación entre el acceso a internet y el desarrollo socioeconómico.

- **Identificación de desigualdades regionales**
  Con el modelo de agrupación, se evaluarán las desigualdades en conectividad y desarrollo entre las regiones de Colombia. Esta etapa permitirá identificar las zonas con mayores limitaciones en acceso a internet y desarrollo socioeconómico, estimando el impacto potencial que tendrían políticas públicas orientadas a mejorar la conectividad en estas áreas, con el fin de reducir las brechas de desigualdad.
- **Validación y ajuste del modelo**
  Para asegurar la precisión del modelo, se validará mediante métricas de rendimiento, como el error medio cuadrático (RMSE) y otras métricas específicas, garantizando que el modelo refleje adecuadamente las características demográficas y socioeconómicas de las regiones estudiadas. Según los resultados obtenidos, se realizarán ajustes necesarios para optimizar el modelo en la agrupación de las regiones.
- **Resultados y recomendaciones**
  Finalmente, se presentará un informe que exponga los hallazgos del modelo de agrupación, identificando las regiones con mayores carencias en conectividad y desarrollo socioeconómico. Basado en estos resultados, se propondrán recomendaciones estratégicas para la formulación de políticas públicas enfocadas en mejorar el acceso a internet en las zonas más vulnerables, promoviendo un desarrollo equitativo en el país.

Descripción Base de Datos Encuesta Nacional de Calidad de Vida - ECV 2023 del DANE:

El proyecto está orientado al diseño de un modelo de agrupación del Desarrollo Socioeconómico de las familias en Colombia a partir de su relación con el acceso a Internet, tomando como base la "Encuesta Nacional de Calidad de Vida - ECV 2023" realizada por el DANE en sus segmentos: Educación; Tecnología; Condiciones de vida de las familias.

Las encuestas de calidad de vida surgen como respuesta a la necesidad de caracterizar la población en los diferentes aspectos involucrados en el bienestar de los hogares. Con el apoyo de la Organización de las Naciones Unidas (ONU),

el Departamento Nacional de Planeación (DNP) y UNICEF, en 1986 se creó en el DANE el proyecto ISPA (indicadores de pobreza absoluta). Su objetivo fue identificar la población en condiciones de pobreza, caracterizarla y ubicar espacialmente.

Las encuestas de calidad de vida son instrumentos que permiten analizar la situación de bienestar de la población investigada. En Colombia, la Encuesta Nacional de Calidad de Vida (ECV) es una operación estadística que el DANE realiza con el objeto de recoger información sobre diferentes aspectos y dimensiones del bienestar y las condiciones de vida de los hogares, incluyendo temas como: el acceso a bienes y servicios públicos, privados o comunales, salud, educación, atención integral de niños y niñas menores de 5 años, entre otros.

La consideración de estos aspectos hace posible realizar análisis a los factores que explican los diferentes niveles de vida existentes en la sociedad. La Encuesta Nacional de Calidad de Vida (ECV) es, además, la fuente de información del cálculo del Índice de Pobreza Multidimensional, IPM.

➢ **Métricas utilizadas:**

Tipo de datos: Encuesta por muestreo (ssd) Unidad de Análisis

Unidades de observación: Corresponden a las viviendas, los hogares y las personas.

Unidades de análisis: También están representadas por las viviendas, los hogares y las personas.

Unidades de muestreo: Existen varias unidades de muestreo dependiendo de la etapa del diseño muestral. Las unidades primarias se relacionan con la primera etapa, las unidades secundarias con la segunda etapa y así sucesivamente.

- Para la Encuesta Nacional de Calidad de Vida (ECV), en particular, se definen 2 etapas con las siguientes unidades de muestreo:
    - Unidades Primarias de Muestreo (UPM): Son todos los municipios del país.
    - Unidades Secundarias de Muestreo (USM): Son conglomerados de 10 viviendas contiguas en promedio, también llamados segmentos o medidas de tamaño (MT), ubicados tanto en la cabecera como en el

resto de cada municipio, con límites fácilmente identificables en los que se encuestan todas las viviendas, hogares y personas.

Versión: Versión 1 (13 de abril de 2023)

Fuente:DANE

- Repositorio de datos oficial del DANE. www.microdatos.dane.go.co

  https://microdatos.dane.gov.co/index.php/catalog/827/get-microdata

- Repositorio de datos propio:

  https://drive.google.com/drive/folders/1n3hU8Ri6n2c2ACj6k88MKLRzsb7JF5JC?usp=drive_link

Segmentos:

- Educación
- Tecnología de información y comunicación
- Condiciones de vida del hogar y tenencia de bienes
- Variables diseño muestra

- **Descripción segmentos "Encuesta Nacional de Calidad de Vida - ECV 2023"**

Con base en los resultados de las ECV 2023 se determinaron tres (3) segmentos tales como Educación; Tecnología y Condiciones de Vida; desde las cuales pueda evaluarse las posibles correlaciones orientadas a explicar y agrupar mediante un modelo el desarrollo socioeconómico de las familias a partir de las condiciones de acceso a internet.

Educación: Esta tabla contiene información sobre educación (principales características educativas de la población de 5 años y más: alfabetismo, asistencia escolar, modalidad de educación, nivel educativo alcanzado y años de estudio; facilidades de acceso a la educación formal en sus diferentes niveles y las razones de inasistencia de la población en edad escolar; cobertura de subsidios, becas y créditos educativos, así como las

entidades que los otorgan; persona con la que permanecen los menores entre 5 y 17 años que no asisten a un establecimiento educativo o después de asistir y los gastos en que incurren los hogares por la asistencia de sus integrantes a establecimientos de educación formal).

Tecnologias de informacion y comunicacion: Esta tabla contiene información sobre las Tecnologías de Información y Comunicación (TIC) para las personas de 5 años y más, con énfasis en la frecuencia de uso del computador, internet y telefonía móvil celular; lugares desde donde se accede a internet, los dispositivos empleados y las actividades realizadas en la web; habilidades en el uso del computador e información sobre los gastos asociados al mismo.

Condiciones de vida del hogar y tenencia de bienes: Esta tabla contiene información sobre las condiciones de vida del hogar y tenencia de bienes (percepción de pobreza, situación económica e inseguridad y la capacidad de los ingresos del hogar para cubrir los gastos mínimos; bienes que posee el hogar, no solo como patrimonio, sino como satisfactores de necesidades y hogares que experimentan inseguridad alimentaria).

- **Descripción de Variables**

La base de datos está dividida en varios segmentos, cada uno representa un aspecto específico del desarrollo socioeconómico de las familias en Colombia. A continuación se enuncian los tres (3) segmentos y algunas variables a partir del diccionario de datos:

1. **Educación**:
    - **Variable:** `P6160`: Indica si la persona sabe leer y escribir (Opciones: 1 = Sí, 2 = No).
    - **Variable:** `P8586`: Muestra si la persona está estudiando actualmente (Opciones: 1 = Sí, 2 = No).
    - **Variable:** `P8587`: Define el nivel educativo más alto alcanzado (Opciones que van desde 1 = Ninguno hasta niveles superiores de educación).

2. **Conectividad a Internet**:
    - Variables de esta dimensión incluirían aspectos como el tipo de conexión a Internet, la frecuencia de uso, niveles de acceso, lo cual es crucial para analizar la relación con el desarrollo socioeconómico.
3. **Situación Económica**:
    - Se incluyen variables como los ingresos familiares, ocupación y otros indicadores financieros.
    - Estas variables cuantifican el estado económico y proporcionan datos numéricos y categóricos que permiten clasificar el bienestar de las familias.

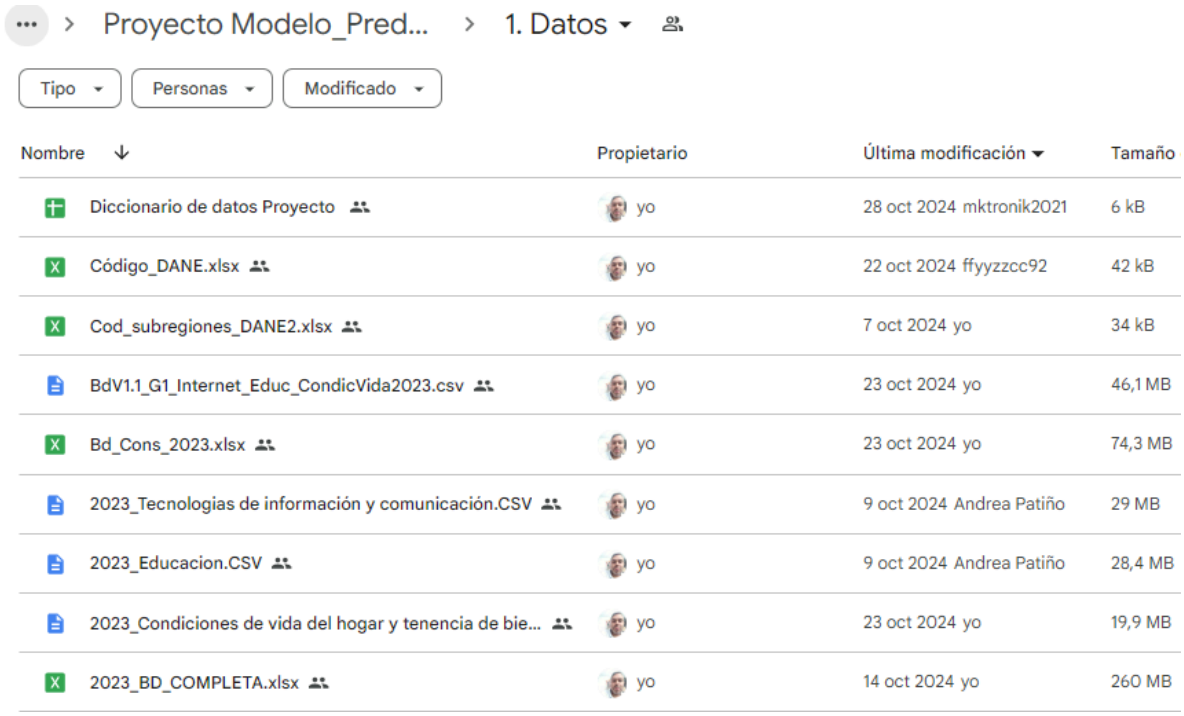
- **Aprestamiento de la base de datos.**

  En tal sentido, se descargaron del repositorio del DANE las bases de datos con respecto a los segmentos anteriores, y sus diccionarios de datos. Este último aspecto permite contar con una mirada particular sobre cada una de las variables, su codificación y opciones de respuestas registradas.

  A partir de la tabla del DANE para municipios y departamentos, se homologaron los nombres de los mismos con base en los códigos registrados en las respuestas en la ECV 2023.

  Finalmente, se procedió a generar una base única consolidada para la cual se construyó un llave a partir de la variable "Directorio" concatenada con la variable "Serie", lo cual corresponde al identificador de cada familia encuestada y a su vez cada miembro de la misma que hizo parte de la encuesta. Así se consolidó la información en una base de datos única, la cual se guardó en formato "CSV" y fue alojada en una carpeta compartida en el Drive en la siguiente ruta:

  https://drive.google.com/drive/folders/1n3hU8Ri6n2c2ACj6k88MKLRzsb7JF5JC?usp=drive_link

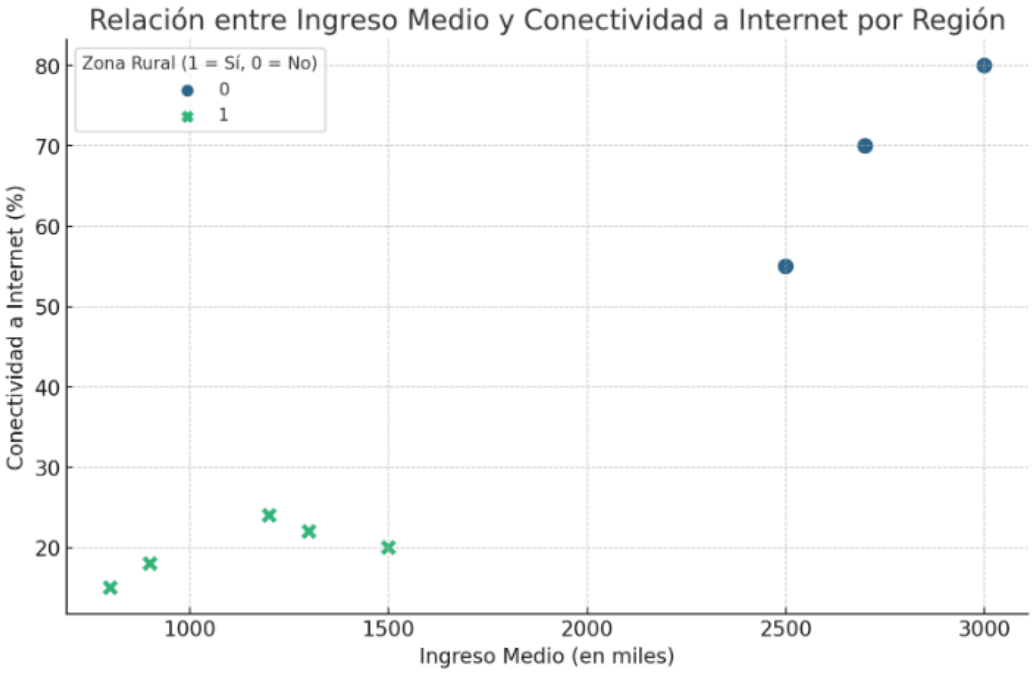


El análisis preliminar de la base de datos permite observar 93 variables con un total de 223.695 registros únicos los cuales corresponden a miembros de familias que respondieron la encuesta. Las variables presentan una codificación propia con base en la estructura particular definida por el DANE.

- **Análisis EDA**

El análisis exploratorio de datos (EDA) del proyecto de conectividad y desarrollo socioeconómico de las familias en Colombia ha arrojado conclusiones clave en torno a la relación entre el acceso a internet y las condiciones socioeconómicas:

1. **Patrones Regionales:** Se observa una fuerte correlación entre la región y las variables de desarrollo económico. Las áreas con menor acceso a internet tienden a coincidir con regiones de menores ingresos y menor desarrollo, lo que sugiere una concentración de la brecha digital en regiones rurales y de bajos recursos.



Aquí tienes la visualización que muestra la relación entre el ingreso medio y la conectividad a Internet por región, destacando las áreas rurales. Se observa que las regiones con menores ingresos tienden a tener menores tasas de conectividad, especialmente en las zonas rurales (indicadas por los puntos de estilo y color específicos). Este patrón refuerza la conclusión sobre la concentración de la brecha digital en áreas de bajos recursos y rurales.

2. **Factores Socioeconómicos y Conectividad:** Variables como el ingreso familiar y el nivel educativo están positivamente asociadas con el acceso a internet. Este patrón es especialmente marcado en zonas urbanas donde el nivel de conectividad es significativamente más alto que en zonas rurales.
3. **Impacto de la Educación en la Conectividad:** La educación superior se relaciona directamente con una mayor conectividad, lo cual indica que en hogares con mayor nivel educativo hay una mayor probabilidad de acceso a internet, reflejando una posible barrera educativa en el acceso digital.
4. **Desigualdades en Conectividad:** Las diferencias en la penetración de internet entre las diferentes regiones del país sugieren que existen desigualdades marcadas, en particular en departamentos alejados de los centros urbanos, lo que implica que la conectividad sigue siendo un reto en zonas de difícil acceso.
5. **Distribución del Acceso a Internet:** La correlación entre variables demográficas, como el tamaño del hogar y la conectividad, indica que los hogares más grandes en zonas rurales tienen menores probabilidades de conectarse a internet, probablemente debido a costos y limitaciones de infraestructura.

Estas observaciones constituyen la base para desarrollar un modelo de agrupación que permita segmentar a las familias según su nivel de desarrollo y conectividad, lo cual facilitará una asignación más eficiente de políticas públicas dirigidas a reducir la brecha digital en las áreas más vulnerables del país.

Para realizar este modelo, se utilizarán algoritmos de agrupación no supervisados como K-means y jerárquicos, evaluando la idoneidad de cada uno en la identificación de estos patrones.

- ## Definición algoritmo K-means

El objetivo principal de K-means es minimizar la varianza dentro de cada clúster al asignar cada punto de datos al grupo con el centroide (promedio) más cercano. Para un conjunto de datos con nnn puntos, el algoritmo busca una partición en KKK clústeres $C1,C2,\ldots,Ck,$ que minimicen la distancia total entre cada punto y su centroide.

Función Objetivo

La función objetivo de K-means es minimizar la suma de las distancias al cuadrado entre cada punto de datos y el centroide de su clúster. Formalmente, la función objetivo JJJ que el algoritmo intenta minimizar se expresa como:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

donde:

- *K* es el número de clústeres,
- *Ci* es el conjunto de puntos asignados al clúster iii,
- *X* representa cada punto de datos en el clúster,
- *μi* es el centroide del clúster *i*,
- ||*x*−*μi*||2 es la distancia euclidiana al cuadrado entre el punto *X* y su centroide *μi*.

Proceso del Algoritmo

El algoritmo K-means se implementa mediante un proceso iterativo, que sigue estos pasos:

I. **Inicialización:** Selecciona *K* puntos aleatorios como centroides iniciales. Estos puntos pueden elegirse aleatoriamente desde el conjunto de datos o mediante métodos como K-means para mejorar la convergencia.
II. **Asignación de Clústeres:** Cada punto de datos se asigna al clúster cuyo centroide esté más cerca, utilizando la distancia euclidiana como medida de similitud.
III. **Recalcular Centroides:** Una vez que todos los puntos han sido asignados a un clúster, se recalcula el centroide de cada clúster como el promedio de los puntos en el grupo.
IV. **Repetición:** Los pasos de asignación y re-cálculos de centroides se repiten hasta que los centroides dejen de cambiar significativamente (convergencia) o se alcance un número máximo de iteraciones.

Elección del Número de Clústeres

Seleccionar el número de clústeres *K* adecuado es clave. Uno de los métodos más comunes es el **Método del Codo (Elbow Method)**, que evalúa el error de agrupamiento (inercia) para distintos valores de KKK. Al graficar el error contra el número de clústeres, el punto donde la disminución del error empieza a ser marginal es el "codo", sugerido como el *K* óptimo.

Ventajas:

- **Simplicidad y eficiencia:** K-means es fácil de implementar y rápido en conjuntos de datos grandes.
- **Interpretabilidad:** Los resultados pueden interpretarse visualmente, especialmente en datos bidimensionales.

Limitaciones:

- **Sensibilidad a la inicialización:** La elección inicial de centroides puede afectar los resultados, a veces resultando en un mínimo local en lugar del mínimo global.
- **Número de clústeres:** El usuario debe especificar *K* de antemano, y un número incorrecto de clústeres puede dar resultados inadecuados.
- **Forma de los clústeres:** K-means agrupa de forma esférica, lo cual puede no ser adecuado si los clústeres tienen formas variadas.

Aplicaciones Prácticas

K-means es ampliamente utilizado para tareas de segmentación en marketing, análisis de comportamiento de usuarios, compresión de imágenes y agrupamiento de documentos, entre otros. En el contexto socioeconómico y de conectividad, como en tu proyecto, K-means permite identificar grupos de familias con características similares, facilitando la comprensión de patrones complejos y ofreciendo una base para intervenciones y políticas personalizadas.

VIII. Modelo.

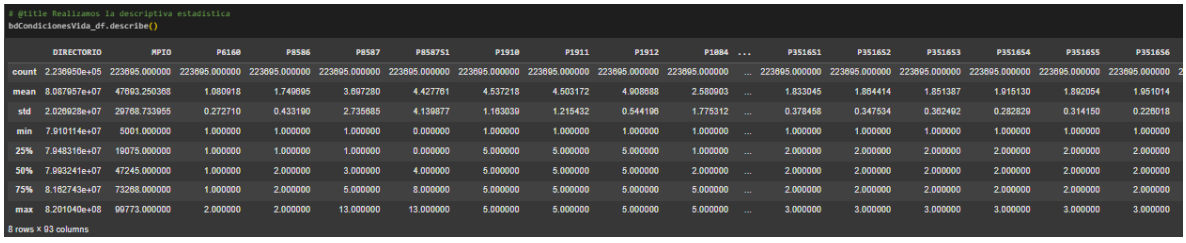
- **Análisis exploratorio.**

El análisis descriptivo constituye uno de los elementos esenciales en el proyecto para entender mejor la estructura de los datos, detectar posibles sesgos y obtener una idea preliminar de las características de cada variable. Esta información es útil para la etapa de análisis exploratorio de datos, especialmente en un modelo de agrupación no supervisado, ya que ayuda a identificar patrones y ajustar la selección de características para el modelo.

Inicialmente se adelantaron algunos análisis con base en estadísticas descriptivas de la base de datos, variables, tipos de datos, calidad de la data. Se ejecutaron funciones orientadas a analizar las características detalladas de la base y determinar métricas de la misma, tales como:

```
# @title Realizamos la descriptiva estadística
bdCondicionesVida_df.describe()
```

Este análisis proporciona información sobre el recuento de datos (`count`), media (`mean`), desviación estándar (`std`), valor mínimo (`min`), percentiles (25%, 50%, 75%), y el valor máximo (`max`) para cada variable del DataFrame**.**



```
# @title Realizamos la descriptiva estadística
bdCondicionesVida_df.describe()
```

	DIRECTORIO	MPIO	P6160	P8586	P8587	P858751	P1910	P1911	P1912	P1084	...	P351651	P351652	P351653	P351654	P351655	P351656
count	2.238950e+05	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	...	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000	223895.000000
mean	8.087957e+07	47693.250368	1.080918	1.749695	3.897280	4.427761	4.537218	4.503172	4.908888	2.580903	...	1.833045	1.884414	1.851387	1.915130	1.892054	1.951014
std	2.026928e+07	29768.733955	0.272710	0.433190	2.735685	4.139877	1.163039	1.215432	0.544196	1.775312	...	0.378458	0.347534	0.362492	0.282829	0.314150	0.226018
min	7.910114e+07	5001.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	7.948316e+07	19075.000000	1.000000	1.000000	1.000000	0.000000	5.000000	5.000000	5.000000	1.000000	...	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
50%	7.993241e+07	47245.000000	1.000000	2.000000	3.000000	4.000000	5.000000	5.000000	5.000000	2.000000	...	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
75%	8.162743e+07	73268.000000	1.000000	2.000000	5.000000	8.000000	5.000000	5.000000	5.000000	5.000000	...	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
max	8.201040e+08	99773.000000	2.000000	2.000000	13.000000	13.000000	5.000000	5.000000	5.000000	5.000000	...	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000

8 rows × 93 columns

Algunos puntos clave de la información en la tabla:

1. Recuento (`count`): Todas las variables tienen el mismo número de datos (223,985), lo cual sugiere que no hay valores faltantes en este conjunto de datos.

2. Media y Mediana: La media y la mediana (representada por el percentil 50%) de cada variable permiten observar las tendencias centrales. Por ejemplo, las variables P687 y P687S1 tienen medias relativamente bajas comparadas con sus máximos, lo que podría indicar valores sesgados hacia la izquierda.
3. Desviación Estándar (`std`): Algunas variables, como P687S1 y P1911, tienen desviaciones estándar más altas, lo que indica mayor variabilidad en esos datos, mientras que otras variables como P31653 tienen una desviación estándar baja, lo que sugiere menor dispersión.
4. Percentiles: Los percentiles proporcionan información adicional sobre la distribución de los datos. Por ejemplo, en varias variables como P1911 y P687S1, el 75% de los datos tiene valores bajos en comparación con sus máximos, lo que podría indicar la presencia de algunos valores atípicos altos.

Posteriormente, se ejecutaron algunas funciones sobre la base de datos orientadas a identificar los municipios, departamentos y regiones del país observables en la data:

1. **Contar registros por departamento**:
    - La primera línea agrupa los datos por departamento y cuenta cuántos registros hay en cada uno. Esto ayuda a saber cuántos datos tienes para cada departamento de Colombia.

```
# Group by 'Departamento' and get the count of records in each department.
department_counts = bdCondicionesVida_df_descritiva.groupby('Departamento')['Departamento'].count()

# Define a dictionary to map departments to regions (you'll need to complete this based on your knowledge of Colombian regions).
region_mapping = {
    'Atlántico': 'Caribe',
    'Bolívar': 'Caribe',
    'La Guajira': 'Caribe',
    'Antioquia': 'Andina',
    'Sucre': 'Caribe',
    'Cundinamarca': 'Andina',
```

2. **Mapeo de departamentos a regiones**:
    - Luego, se define un diccionario (`region_mapping`) que relaciona cada departamento de Colombia con su respectiva región (Andina, Caribe, Pacífico, etc.). Este mapeo permite analizar los datos no solo a nivel de departamento, sino también a nivel de región.

```
# Add a 'Region' column to your dataframe using the mapping.
bdCondicionesVida_df_descritiva['Region'] = bdCondicionesVida_df_descritiva['Departamento'].map(region_mapping)

# Now you can group by 'Region' to analyze data by region.
region_counts = bdCondicionesVida_df_descritiva.groupby('Region')['Region'].count()

# Display the counts of records per region.
region_counts
```

3. Agregar una columna de "Región":

- Usando el mapeo, el código añade una columna nueva llamada `Region` en el conjunto de datos. Esta columna indica la región de cada departamento.

4. Contar registros por región:

- Finalmente, el código agrupa los datos por región y cuenta cuántos registros hay en cada una, permitiendo ver cuántos datos hay en total en cada región de Colombia.

Este código es útil para el proyecto y el análisis del impacto de Internet en el nivel socioeconómico en Colombia, teniendo en cuenta:

1. **Análisis Regional**:
    - Al agregar la región, se observan patrones a nivel regional, no solo por departamento. Esto puede ayudar a identificar diferencias significativas entre regiones, como cuál región tiene más acceso a Internet o mejores condiciones socioeconómicas.
2. **Visualización Simplificada**:
    - Agrupar por región permite visualizar los datos de manera más resumida, lo cual es útil para análisis y gráficos comparativos. En lugar de ver información de 32 departamentos, se resume en 5 o 6 regiones.
3. **Identificación de Regiones con Mayor Necesidad**:
    - Este análisis puede mostrar si ciertas regiones están más rezagadas en términos de acceso a Internet o condiciones de vida, permitiendo enfocarse en las áreas que necesitan más apoyo.

Adicionalmente, se utilizó un código para crear un gráfico de caja (boxplot) que compara los departamentos en cuanto a una variable, en este caso que indica si las personas saben leer y escribir y así visualizar cómo la educación varía entre

departamentos en Colombia, lo cual es fundamental para entender el contexto de acceso a Internet y su relación con el nivel educativo y socioeconómico.

1. **Definir el tamaño del gráfico**:

- La primera línea ajusta el tamaño del gráfico, para que sea más fácil de visualizar, es decir, este gráfico tiene un ancho de 10 pulgadas y una altura de 6 pulgadas, hace que sea más legible y tenga el tamaño adecuado para visualizarse bien especialmente cuando hay varios elementos o etiquetas.

```
# @title Grafica de Boxplot composición de la base de datos
plt.figure(figsize=(10, 6))
# Seleccionar los departamentos con mayor cantidad de registros: Variables P6160  P8586 P8587 P8587S1
Cantidad_de_departamentos = 10
nombres_Departamentos = bdCondicionesVida_df_descritiva['Departamento'].value_counts().iloc[:Cantidad_de_departamentos].index.tolist()
```

2. **Seleccionar los departamentos con más registros**:

- Luego, selecciona los 10 departamentos con la mayor cantidad de datos en la base de datos. Esto permite centrarse en las zonas con más información, lo cual puede hacer el análisis más representativo.

```
sns.boxplot(x='Departamento', y='P6160', data=bdCondicionesVida_df_descritiva[bdCondicionesVida_df_descritiva['Departamento'].isin(nombres_Departamentos)])
plt.title('Boxplot de Departamentos vs. Sabe leer y escribir P6160')
plt.xlabel('Departamento')
plt.ylabel('Educación')
plt.xticks(rotation=90)
plt.show()
```

3. **Crear el boxplot**:
- El código usa `seaborn` para crear un gráfico de caja que muestra la distribución de la variable `P6160` (posiblemente, "saber leer y escribir") en cada uno de los 10 departamentos seleccionados.
- En el eje x están los nombres de los departamentos, y en el eje y, la variable de educación (`P6160`).

4. **Etiquetas y rotación**:
- El título y las etiquetas de los ejes se agregan para hacer el gráfico más claro. Además, el nombre de cada departamento en el eje x se rota para que sea más fácil de leer.

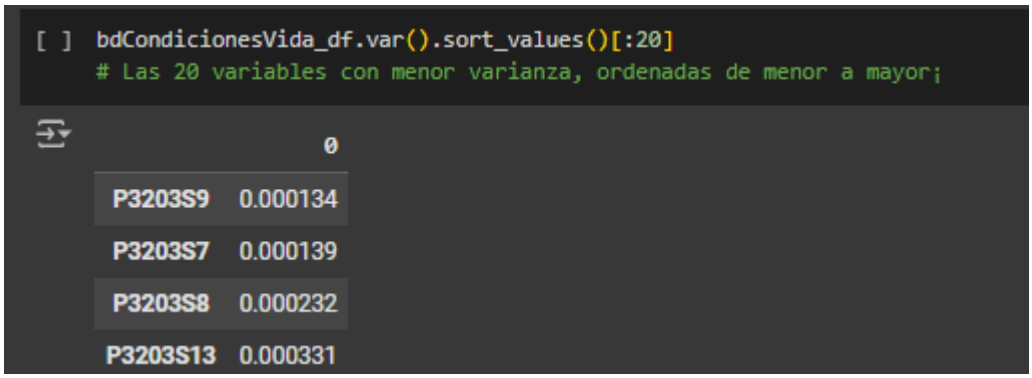
Este gráfico puede aportar mucho al proyecto sobre el impacto de Internet en el nivel socioeconómico en Colombia.

1. **Comparación de Alfabetización entre Departamentos**:
    - Este gráfico permite observar la distribución de la alfabetización (saber leer y escribir) en distintos departamentos. Así se pueden identificar diferencias o desigualdades en la educación básica, un aspecto importante al analizar el nivel socioeconómico.
2. **Detección de Departamentos con Mayores Necesidades**:
    - Si algunos departamentos muestran menor alfabetización o más variación, esto podría indicar una mayor necesidad de recursos o programas de alfabetización. Esto es relevante para ver cómo el acceso a Internet podría ayudar en estas zonas.
3. **Visualización de Datos y Análisis Exploratorio**:
    - Este tipo de gráfico ayuda a entender rápidamente la estructura de los datos y es una buena herramienta de análisis exploratorio. Da

una idea de la calidad de los datos en cada departamento y muestra si hay valores atípicos (puntos alejados de los demás).

El siguiente código ejecutado tiene como función identificar las variables menos informativas, permitiendo tomar decisiones sobre cuáles incluir o excluir en el análisis.

Esto optimiza el proyecto y permite trabajar en las variables que podrían explicar mejor la relación entre el acceso a Internet y el nivel socioeconómico en Colombia.



Calcular la varianza de cada columna:

- `bdCondicionesVida_df.var()` calcula la varianza de cada variable (columna) en el DataFrame `bdCondicionesVida_df`.
- La varianza mide cuánto varían los datos de una variable con respecto a su media. Si la varianza es baja, significa que los datos están muy cerca de la media y, por lo tanto, no cambian mucho.

Ordenar las variables por varianza:

- `.sort_values()` organiza estas varianzas en orden ascendente (de menor a mayor).

Seleccionar las primeras 20 variables:

- `[:20]` toma solo las primeras 20 variables de esta lista, que son las 20 variables con menor varianza.

Para el proyecto, esta información es útil por varias razones:

1. **Identificación de Variables Poco Informativas**:
    - Las variables con baja varianza suelen ser poco informativas, ya que casi todos los valores son iguales o muy similares. Esto significa que probablemente no contribuyen mucho al análisis del impacto del Internet en el nivel socioeconómico. Así se puede decidir si eliminarlas o ignorarlas en el análisis para simplificar el modelo y enfocarse en las variables más relevantes.
2. **Optimización del Modelo**:
    - Reducir el número de variables irrelevantes o redundantes puede hacer que el análisis sea más rápido y preciso. Es mejor centrarse en variables con mayor variabilidad, ya que suelen tener más información valiosa para el modelo.
3. **Posible Revisión de Datos**:
    - Estas variables de baja varianza pueden indicar que ciertos datos se recopilaron de manera homogénea o que son constantes en el tiempo. Esto puede llevar a revisar si estos datos son relevantes para el análisis o si deberían ajustarse.

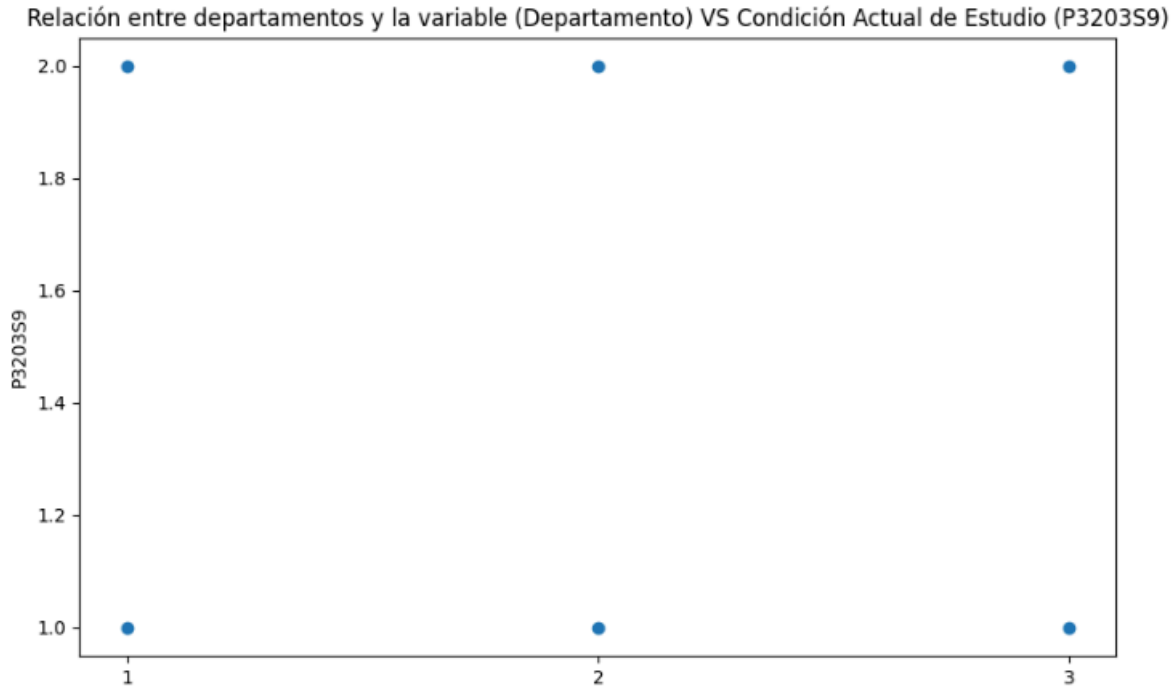
El siguiente código es útil para crear una visualización sencilla de datos que permite entender la relación entre variables, lo cual es muy valioso en análisis exploratorio y visualización de datos. En este caso se indica si los hijos fueron retirados de sus estudios esta situación puede variar entre departamentos en Colombia, lo cual es fundamental para entender el contexto de acceso a Internet y su relación con el nivel educativo y socioeconómico en tu proyecto.

```python
# Importa las librerías necesarias

import pandas as pd
import matplotlib.pyplot as plt

# Supongamos que bdCondicionesVida_df_descriptiva ya está definido
# Ejemplo de cómo podría ser tu DataFrame
bdCondicionesVida_df_descriptiva = pd.DataFrame({
    'Departamento': [1, 2, 3, 1, 2, 3],
    'P3203S9': [1, 2, 1, 2, 1, 2]
})

plt.figure(figsize=(10, 6))
plt.scatter(
    bdCondicionesVida_df_descriptiva['Departamento'].astype(str),  # Convertimos a string
    bdCondicionesVida_df_descriptiva['P3203S9'].astype(float)      # Aseguramos que esta columna sea numérica
)
plt.title('Relación entre departamentos y la variable (Departamento) VS Condición Actual de Estudio (P3203S9)')
plt.xlabel('Departamento')
plt.ylabel('P3203S9')
plt.show()
```



Creación de Datos de Ejemplo en un DataFrame:

- Se define un `DataFrame` (tabla de datos) llamado `bdCondicionesVida_df_descriptiva`. Este `DataFrame` tiene dos columnas:
    - **"Departamento"**: Contiene valores 1, 2 o 3, que podrían representar distintos departamentos o regiones.

    - **"P3203S9"**: Contiene valores 1 o 2, que podrían indicar diferentes estados o condiciones de estudio para cada registro.
- Esto es un ejemplo de cómo podrían ser los datos en el `DataFrame` para fines de demostración.

Creación de un Gráfico de Dispersión:

- Se utiliza `matplotlib` para generar un gráfico de dispersión.
- En el gráfico:
    - **Eje X**: Muestra la variable "Departamento".
    - **Eje Y**: Muestra la variable "P3203S9" (Condición Actual de Estudio).
- Cada punto en el gráfico representa un par de valores para "Departamento" y "P3203S9".

- **Visualización de la Relación entre Variables**: Este gráfico permite observar si existe alguna relación visual entre los departamentos y la condición actual de estudio de los individuos. En este caso, se puede ver que hay puntos en ciertos departamentos y en determinados valores de condición.

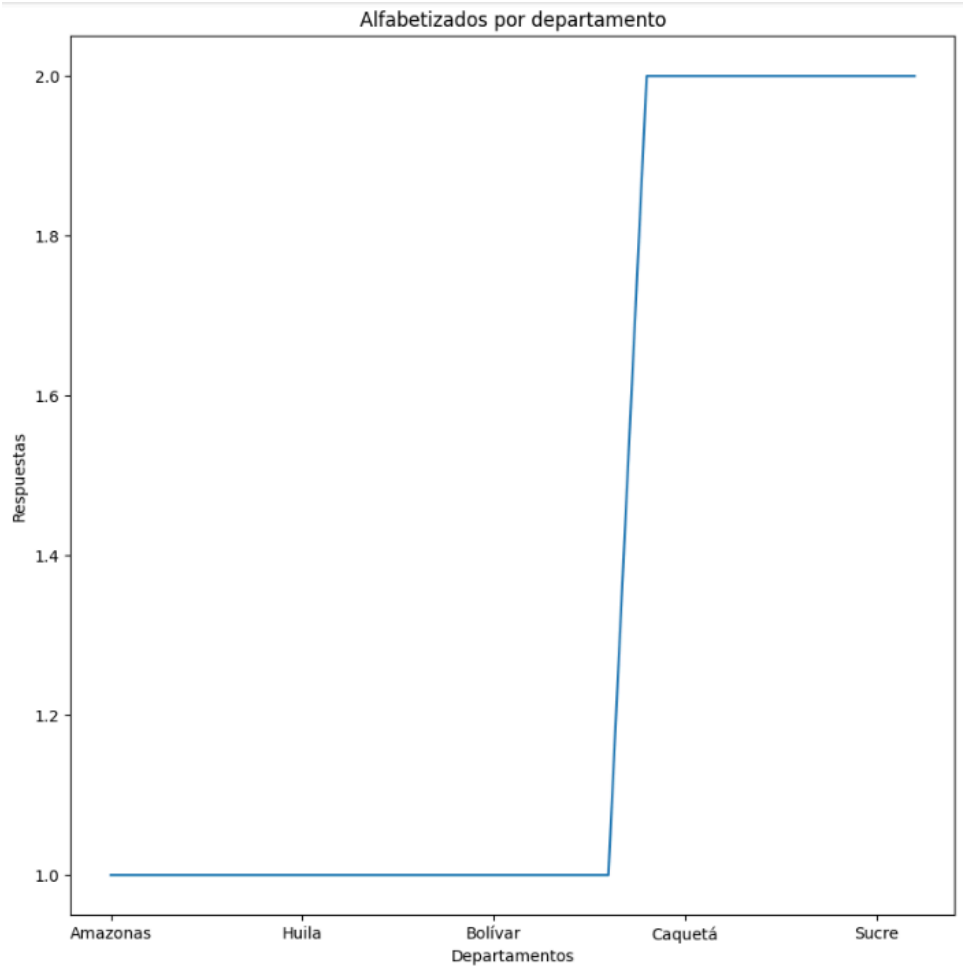
- **Identificación de Tendencias o Agrupaciones**: Aunque los datos son muy limitados (sólo algunos puntos), este tipo de visualización sería útil para proyectos más grandes con más datos, ya que puede ayudar a identificar patrones o concentraciones en ciertas categorías.

- **Verificación de Datos**: La gráfica es también una forma rápida de verificar si los datos están en el formato correcto (numérico en "P3203S9", categórico en "Departamento") y si tienen la distribución esperada.

El siguiente código es útil para generar una visualización rápida y sencilla que ayuda a entender la distribución de la alfabetización en diferentes departamentos, lo cual puede ser valioso en proyectos que busquen mejorar condiciones educativas o evaluar políticas públicas.

Este código en Python utiliza un `DataFrame` llamado `bdCondicionesVida_df_descriptiva` para hacer una gráfica de líneas. La función principal del código es contar cuántas personas alfabetizadas hay en cada departamento y luego graficar esta información.

```
#@title Respuestas de Alfabetizados por departamento
plt.figure(figsize=(10, 10))
bdCondicionesVida_df_descritiva[bdCondicionesVida_df_descritiva.P3203S9==1].groupby(by=['Departamento'])['P3203S9'].count().sort_values().plot(kind='line')
plt.title('Alfabetizados por departamento')
plt.xlabel('Departamentos')
plt.ylabel('Respuestas')
plt.show()
```



Filtrar Datos:

- La línea `bdCondicionesVida_df_descriptiva[bdCondicionesVida_df_descriptiva.P3203S9==1]` filtra el `DataFrame` para obtener solo aquellos registros en los que la columna `P3203S9` tiene el valor `1`, lo cual probablemente indica que esas personas están alfabetizadas.

Agrupar y Contar Respuestas:

- La parte `.groupby(by=['Departamento'])['P3203S9'].count()` agrupa los datos filtrados por el valor de la columna "Departamento" y cuenta el número de respuestas (personas alfabetizadas) en cada departamento.

Ordenar y Graficar:

- `.sort_values().plot(kind='line')` ordena los departamentos de menor a mayor cantidad de personas alfabetizadas y genera una gráfica de línea que muestra esta información.

Configuración de la Gráfica:

- `plt.figure(figsize=(10, 10))` establece el tamaño de la figura.
- `plt.title`, `plt.xlabel`, y `plt.ylabel` añaden un título y etiquetas a los ejes de la gráfica.
- `plt.show()` muestra la gráfica.

Visualización de la Alfabetización por Departamento: Este código permite visualizar la cantidad de personas alfabetizadas en cada departamento. Esto es útil para entender cómo se distribuye la alfabetización en distintas regiones.

Identificación de Disparidades: La gráfica de líneas facilita identificar departamentos con altas o bajas cantidades de personas alfabetizadas, lo cual puede ayudar a detectar regiones que podrían necesitar más apoyo educativo.

Información para la Toma de Decisiones: Estos datos pueden ser clave para organismos que deseen enfocar esfuerzos en mejorar la alfabetización en departamentos con menores cifras, permitiendo una asignación de recursos más informada.

El siguiente código en Python que se utilizó visualiza una **matriz de correlación** entre variables numéricas. La matriz de correlación muestra cómo se relacionan entre sí diferentes variables, ayudando a entender si hay alguna relación lineal entre ellas (positiva, negativa o ninguna).

```
# @title Visualización de la correlación entre variables numéricas
import matplotlib.pyplot as plt
plt.figure(figsize=(20, 20))
sns.heatmap(conf_matrix, annot=False, cmap='coolwarm', fmt=".2f")
plt.title('Matriz de Correlación')
plt.show()
```

Importar la Biblioteca para Graficar:

- `import matplotlib.pyplot as plt`: Importa la biblioteca `matplotlib.pyplot` y le da el alias `plt`, lo que permite crear visualizaciones en Python.

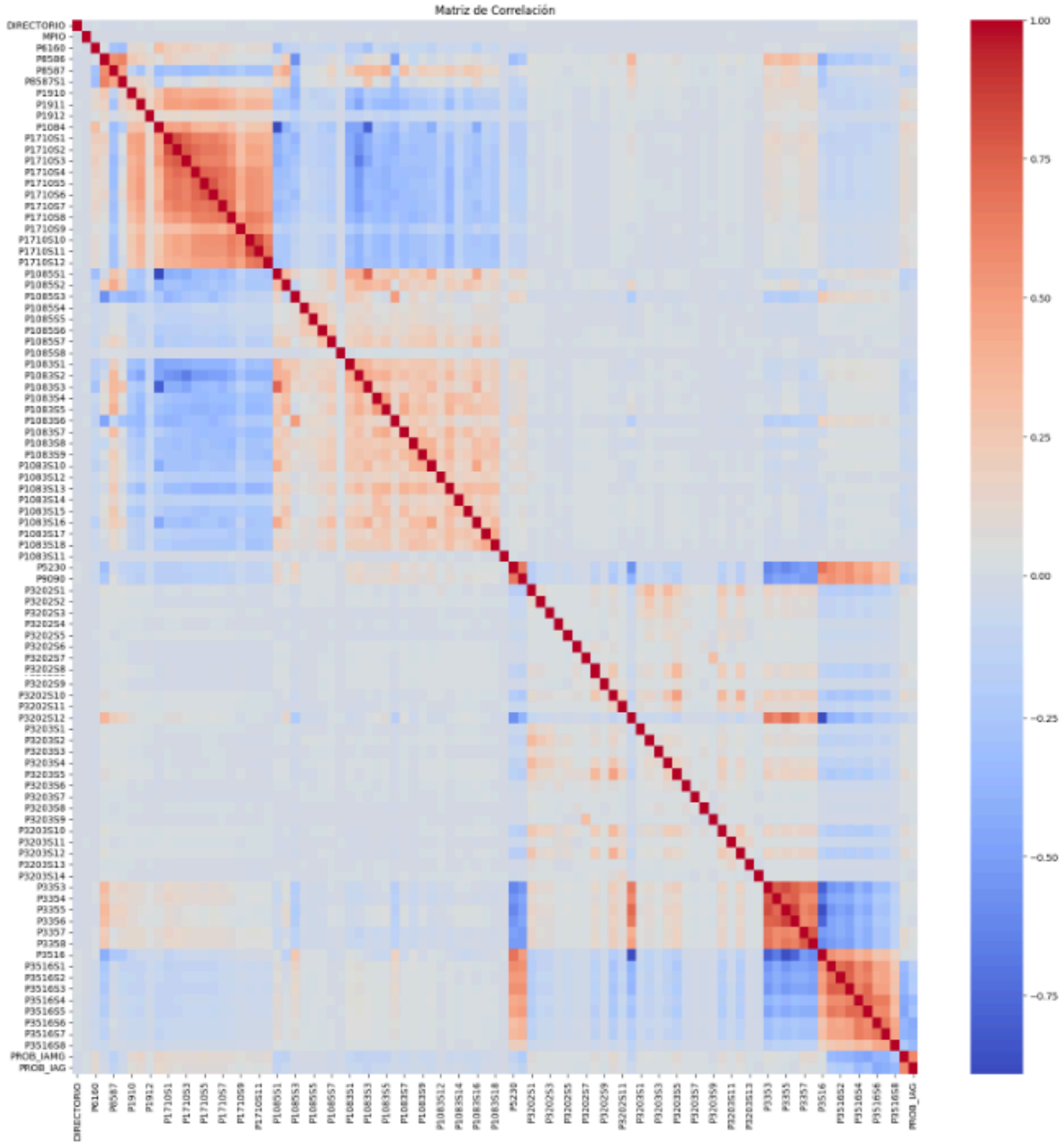
Crear una Figura:

- `plt.figure(figsize=(20, 20))`: Establece el tamaño de la figura en 20x20 pulgadas. Esto hace que la gráfica sea bastante grande, lo cual es útil si hay muchas variables en la matriz.

Generar el Mapa de Calor (Heatmap):

- `sns.heatmap(conf_matrix, annot=False, cmap='coolwarm', fmt=".2f")`: Usa `seaborn` (sns) para crear un mapa de calor con la matriz de correlación (`conf_matrix`). Aquí se especifican varios parámetros:
  - `conf_matrix`: Es la matriz de correlación (debe haberse calculado antes y almacenado en esta variable).
  - `annot=False`: No muestra los valores numéricos en cada celda (si quisieras ver los valores numéricos, podrías cambiar esto a `True`).
  - `cmap='coolwarm'`: Es el color que va desde tonos fríos a cálidos, lo cual permite distinguir fácilmente entre correlaciones positivas y negativas.

- `fmt=".2f"`: Formatea los números dentro del heatmap para mostrar solo dos decimales, en caso de que se muestre `annot=True`.



Visualización de Correlaciones: Esta matriz es útil para identificar rápidamente las relaciones entre variables numéricas en el conjunto de datos. Las correlaciones positivas (cercanas a 1) indican que cuando una variable aumenta, la otra también tiende a aumentar, mientras que correlaciones negativas (cercanas a -1) indican que cuando una variable aumenta, la otra tiende a disminuir.

Selección de Variables: La matriz de correlación permite ver si hay variables que están muy correlacionadas entre sí, lo cual podría indicar redundancia. Esto es útil para simplificar modelos y eliminar variables que aportan información duplicada.

Interpretación de Datos: Al analizar esta matriz, es posible entender mejor la estructura de los datos, lo cual facilita la interpretación de patrones o comportamientos de las variables en el proyecto.

- **Resultados y conclusiones**

1. **Análisis de correlación entre las variables numéricas.** La correlación es una medida que indica la relación entre dos variables y puede ayudar a identificar características que podrían ser relevantes para el modelo.

   En el contexto de este proyecto, que utiliza un modelo de agrupación no supervisado, el análisis de correlación ayuda a identificar las variables que están fuertemente relacionadas entre sí, teniendo en cuenta que:

   Evita redundancia: En un modelo de agrupación, variables muy correlacionadas pueden aportar información repetida, lo cual puede distorsionar la agrupación. Con esta matriz de correlación, es posible detectar estas relaciones y considerar eliminar o combinar algunas variables altamente correlacionadas.

   Mejora la interpretabilidad: Al comprender qué variables tienen una relación más fuerte, el análisis de agrupación puede centrarse en aquellas características clave que representan la variabilidad entre grupos de manera más efectiva.

   Optimiza el rendimiento: Reducir la cantidad de variables redundantes simplifica el modelo, haciéndolo más eficiente y reduciendo el tiempo de procesamiento.

   ### Explicación del código:

   Selección de columnas numéricas:

   ```
   numeric_cols = bdCondicionesVida_df_descritiva.select_dtypes(include=['number'])
   ```

   ➢ Esta línea selecciona todas las columnas numéricas del DataFrame `bdCondicionesVida_df_descritiva`, almacenándolas en

`numeric_cols`. Este paso es útil porque la matriz de correlación sólo puede calcularse entre variables numéricas.

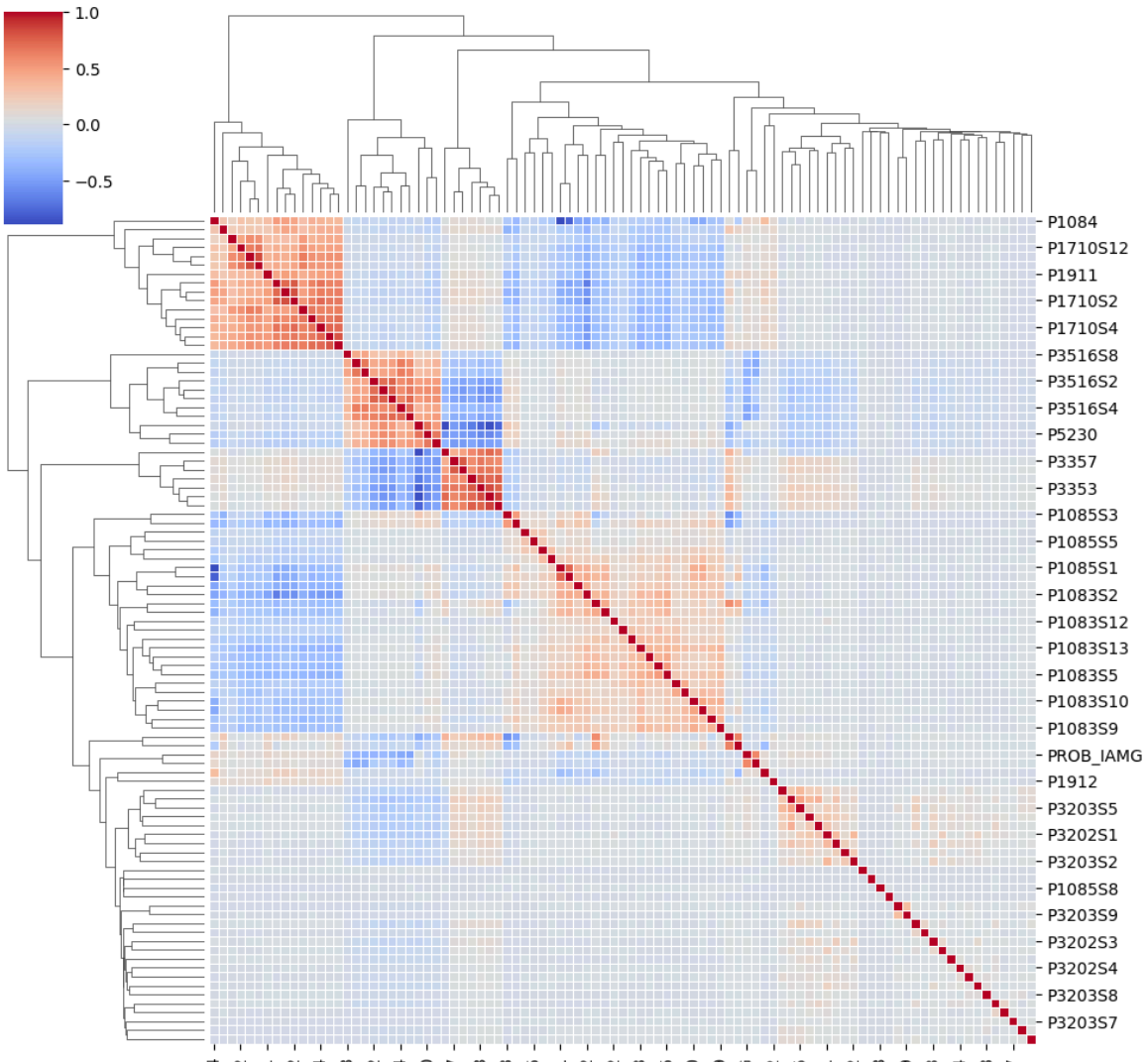
Cálculo de la matriz de correlación:

```
conf_matrix = numeric_cols.corr()
```

➢ Esta línea calcula la **matriz de correlación** de las columnas numéricas, utilizando el método `.corr()`. El resultado es una matriz cuadrada donde cada elemento representa el coeficiente de correlación entre dos variables.

2. **Representación de clúster jerárquico (o clustermap)**

Con el objetivo de visualizar la correlación entre las variables numéricas y los patrones entre las mismas, se genera un dendrograma con base en clúster jerárquicos sobre la matriz inicial:



El dendrograma muestra cómo las variables se agrupan según su nivel de correlación. Esto ayuda a identificar variables que podrían estar representando la misma información, lo cual es relevante para optimizar el modelo.

Esto permite analizar cómo las variables se agrupan en función de su similitud o correlación, lo cual es relevante para entender la estructura de los datos y seleccionar variables, lo cual es esencial para seleccionar variables y optimizar el rendimiento de un modelo de agrupación.

3. **Partición de datos y transformación de variables.**
   - **Escalamiento de los datos.**

```
df_scaled = bdCondicionesVida_df_descriptiva[row_labels[:33]]
```

Se seleccionan las primeras 33 características del conjunto `bdCondicionesVida_df_descriptiva`, las cuales están ordenadas por su correlación agrupada (derivada de los clústeres en el análisis previo). `df_scaled` contiene solo estas variables, filtrando aquellas que podrían tener menor relevancia o estar muy correlacionadas con otras.

Al seleccionar solo las características más relevantes y agrupadas por correlación, se disminuye la complejidad del modelo, lo que puede llevar a una mejor eficiencia y rendimiento.

- **División de los Datos en Conjuntos de Entrenamiento y Prueba**:

```
from sklearn.model_selection import train_test_split

X_train, X_test = train_test_split(df_scaled, test_size=0.3, random_state=42)
```

Se utiliza ***train_test_split*** de sklearn para dividir df_scaled en dos conjuntos: entrenamiento y prueba, con un 70% para entrenamiento y 30% para prueba. El parámetro random_state=42 asegura que esta división sea

reproducible, lo cual es importante para obtener resultados consistentes en experimentos.

La división en entrenamiento y prueba garantiza que el modelo se entrene en un subconjunto de los datos y se evalúe en datos nuevos, lo cual es fundamental para medir su rendimiento real y evitar el sobreajuste (overfitting).

4. **Análisis de Componentes Principales (PCA).**

El PCA es una técnica efectiva para reducir la dimensionalidad de los datos, lo cual puede mejorar el rendimiento de los modelos al eliminar redundancias y reducir el ruido, disminuir el tiempo de entrenamiento y el consumo de recursos y evitar el sobreajuste, especialmente en conjuntos de datos con muchas características.

- **Configuración del Número de Componentes**:
  `n_components = 2  # Parámetro ajustable con base en las necesidade del modelo.`

Define el número de componentes principales que deseas mantener. En este caso, se ha configurado en 2, pero puedes ajustar este valor según los resultados deseados o la variabilidad que quieras preservar.

- **Creación y Ajuste del Modelo PCA**:
  `pca = PCA(n_components=n_components)`

`X_train_pca = pca.fit_transform(X_train)`

Se crea un objeto `PCA` con el número especificado de componentes. Luego, el método `fit_transform` se aplica a los datos de entrenamiento `X_train`, transformándolos en una nueva representación de menor dimensionalidad, `X_train_pca`.

En este caso, después de aplicar el PCA, X_train_pca contiene los datos transformados con una dimensionalidad reducida a solo los dos componentes principales especificados, preservando la mayor parte de la variabilidad relevante en los datos.

- **Proporción de varianza explicada.**

El resultado de la proporción de varianza explicada por cada componente y la acumulada indica cuánta información de los datos originales se ha capturado en los componentes seleccionados, que para el modelo inicial fueron:

```
Proporción de varianza explicada por cada componente: [0.33927004 0.26291366]
Varianza explicada acumulada: [0.33927004 0.6021837 ]
```

Proporción de varianza explicada por cada componente:

- El primer componente principal explica el 33.93% de la variabilidad de los datos.
- El segundo componente principal explica el 26.29% de la variabilidad.

Varianza explicada acumulada:

- La varianza acumulada después de dos componentes es del 60.22%.
- Esto significa que al reducir la dimensionalidad a dos componentes, se está reteniendo el 60.22% de la información original de los datos.

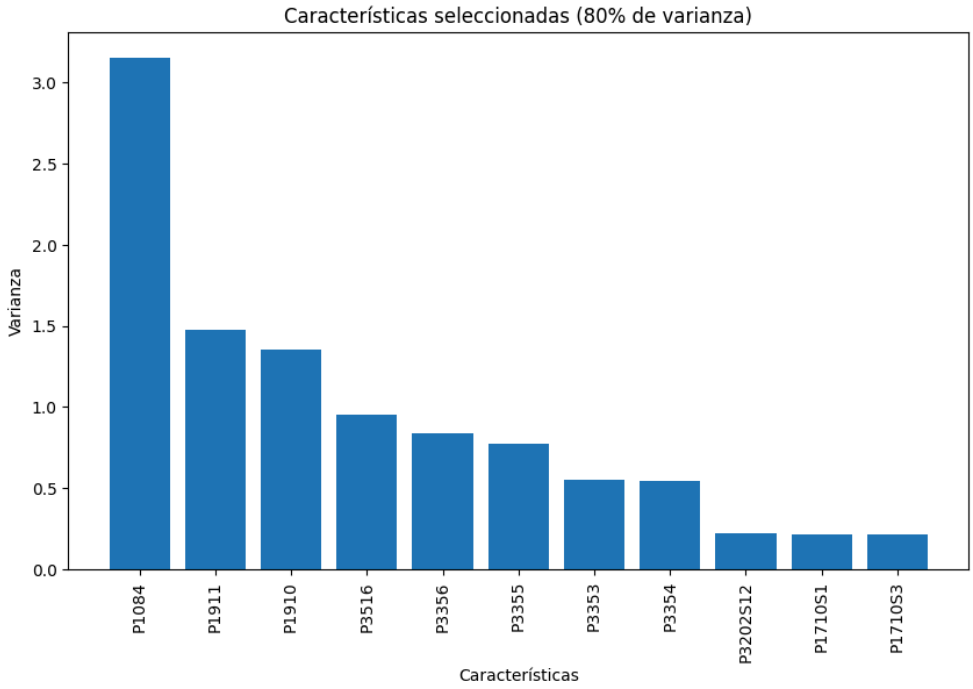
Interpretación de cara al modelo:

1. **Conservación de Información:**
   - Un valor de 60.22% de varianza explicada acumulada significa que los dos componentes principales retienen un porcentaje significativo de la información original. Sin embargo, también implica que un 39.78% de la variabilidad de los datos se está perdiendo al reducirlos a dos dimensiones. Dependiendo del caso, esta pérdida puede ser aceptable o puede indicar que sería necesario agregar más componentes.

2. **Reducción de la Dimensionalidad:**
    - Mantener sólo dos componentes puede simplificar mucho el modelo, lo que podría ayudar en términos de tiempo de procesamiento y disminuir el riesgo de sobreajuste. Sin embargo, si el modelo necesita capturar más detalles, se podría considerar aumentar el número de componentes para captar una mayor parte de la varianza.

3. **Impacto en la precisión del modelo:**
    - La cantidad de varianza explicada acumulada que es suficiente depende del contexto y de la sensibilidad del modelo a la pérdida de información.
    - Con el 60.22%de los datos explicados, es posible que el modelo no capte todas las variaciones y patrones, especialmente si las dimensiones eliminadas contienen información relevante.
    - Aumentar los componentes hasta capturar un porcentaje de varianza superior, puede ser una alternativa para observar el comportamiento de los patrones, la varianza de los mismos y posiblemente mejorar el rendimiento.
    - Alternativamente, si el modelo funciona bien con este nivel de variabilidad retenida, podría mantenerse en dos componentes para maximizar la simplicidad y la interpretabilidad.

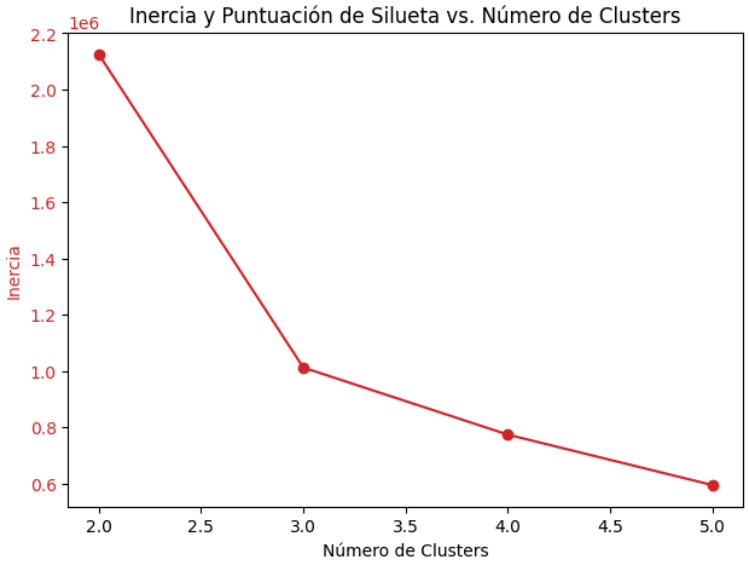
4. **Selección características agrupadas en las varianzas acumuladas hasta el 80%.**



La gráfica muestra las características seleccionadas que explican el 80% de la varianza en el modelo, indicando que estas variables son las más relevantes para capturar la diversidad en los datos relacionados con el desarrollo socioeconómico y la conectividad de las familias en Colombia, teniendo en cuenta:

- La variable P1084 (Frecuencia de uso del internet) tiene una varianza significativamente más alta que las demás, lo que sugiere que aporta una mayor capacidad de diferenciación entre familias en términos de desarrollo socioeconómico y la conectividad.
- Al enfocarse en las variables que explican la mayor parte de la varianza, el modelo puede agrupar de manera más precisa a las familias en diferentes clusters o grupos, basados en características clave.
- Al reducir las variables a solo aquellas que explican el 80% de la varianza, el modelo se vuelve más interpretativo, lo que facilita la identificación de patrones de conectividad y desarrollo socioeconómico en los distintos grupos de familias.

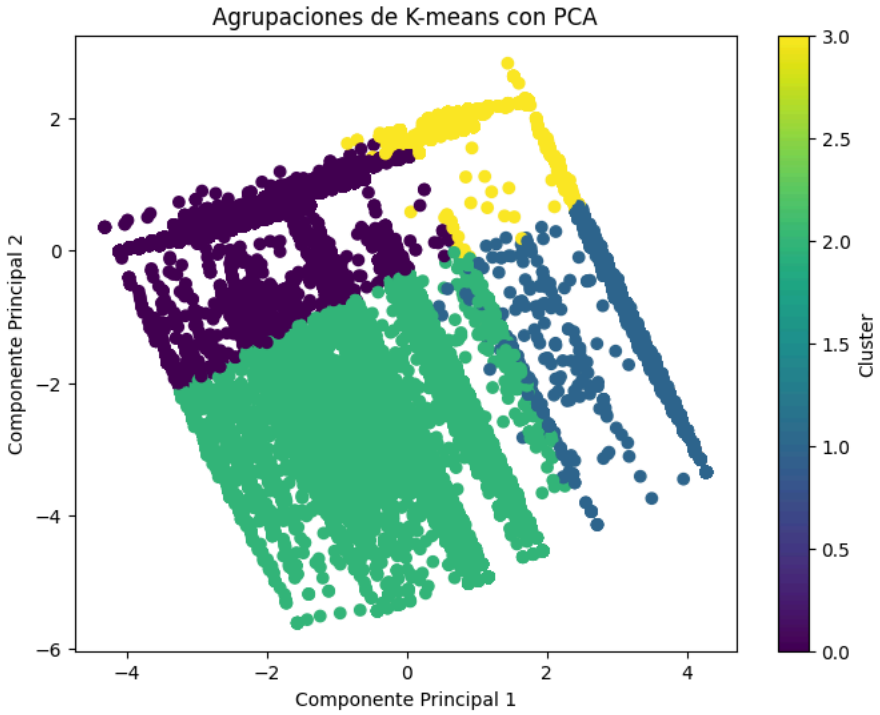
5. **Gráfica de inercia Vs Número de Cluster.**



La gráfica que compartismuestra la inercia y la puntuación de silueta en función del número de clusters en un modelo de K-means. Estos dos indicadores son útiles para determinar el número óptimo de clusters en el análisis de agrupamiento, encontrando que:

En la gráfica, la inercia disminuye significativamente al aumentar el número de clusters de 2 a 3 y luego de 3 a 4, pero la reducción es menor a partir de 4 clusters. Este "punto de codo" indica que 3 o 4 clusters podrían ser opciones adecuadas, ya que más clusters no resultan en una mejora significativa de la inercia.

6. Agrupaciones de K-Means con PCA



La gráfica permite observar los resultados del modelo de K-means aplicado a los datos transformados mediante Análisis de Componentes Principales (PCA), visualizando las agrupaciones en función de los dos primeros componentes principales.

Se observan claramente cuatro clusters distintos (etiquetados de 0 a 3 en la barra de color). La separación entre clusters sugiere que el modelo K-means logró identificar subgrupos con patrones diferentes en los datos, probablemente vinculados a combinaciones particulares de conectividad e indicadores socioeconómicos. El cluster en verde (etiqueta 1) es el más disperso, mientras que el cluster en amarillo (etiqueta 3) parece concentrarse en un área específica.

a. **Cluster de alta frecuencia de uso de Internet.** Los hogares en este cluster presentan un uso de Internet frecuente, lo que sugiere que la conectividad es una parte fundamental de su vida cotidiana. Este grupo tiende a ubicarse en áreas urbanas con mayores niveles de ingreso y acceso a infraestructura digital.

b. **Cluster de uso intermitente o moderado de Internet.** Este cluster representa hogares que tienen acceso a Internet, pero su uso es intermitente o moderado. Esto podría deberse a limitaciones económicas o de infraestructura que dificultan el acceso regular. Los hogares en este grupo suelen estar en áreas semiurbanas o rurales, donde la conectividad no es tan accesible.

c. **Cluster de baja o nula frecuencia de uso de Internet.** Los hogares en este cluster muestran una frecuencia de uso de Internet muy baja o inexistente, lo que generalmente corresponde a zonas rurales o de bajos ingresos donde la infraestructura de conectividad es escasa o inexistente. La falta de acceso frecuente limita su desarrollo socioeconómico, restringiendo el acceso a servicios digitales esenciales como la educación, el empleo y el acceso a información actualizada.

Conclusiones.

1. El análisis exploratorio de datos (EDA) permite observar que los hogares con un mayor nivel educativo tienen más acceso a internet, sugiriendo una barrera educativa en el acceso digital.
2. El modelo de agrupación no supervisado (K-means y jerárquico) permitió identificar patrones diferenciados de desarrollo socioeconómico y conectividad a internet en distintas regiones de Colombia.
3. El uso de técnicas de reducción de dimensionalidad y selección de características permitió que el modelo se centrara en las variables más significativas, mejorando la interpretación de los resultados.
4. Los modelos utilizados en diferentes variaciones confluyen al final en identificación de tres (3) clusters definidos, en su relación a la variables P1084 (Frecuencia de uso del internet):
    - **Cluster de alta frecuencia de uso de Internet.** La alta frecuencia de uso está asociada a mejores oportunidades educativas y laborales, lo que refuerza la relación entre desarrollo socioeconómico y conectividad constante.
    - **Cluster de uso intermitente o moderado de Internet.** Aunque estos hogares utilizan Internet, es probable que el acceso limitado impacte su capacidad para aprovechar oportunidades de educación en línea o de trabajo remoto.
    - **Cluster de baja o nula frecuencia de uso de Internet.** Este patrón indica una clara necesidad de políticas de inclusión digital dirigidas a reducir la brecha de conectividad en las regiones más vulnerables.
5. La identificación de clusters específicos permite una intervención más efectiva en la posible intervención a través de política pública, propendiendo mejorar las condiciones de vida en las regiones de menor conectividad y promoviendo una Colombia más conectada y equitativa.

X. Anexos.

A. Presupuesto

COSTOS DEL PROYECTO	
ITEM	**VALOR**
LAPICERO	$2.000
INTERNET	$129.000
ENERGIA	$80.000
ASESORIA TECNICA DE EXPERTOS	$3.200.000
ANÁLISIS Y DISEÑO MODELO	$4.000.000
TOTAL	**$7.411.000**

B. Diagrama de flujo del diccionario de datos.

C. Cuaderno en Colab (Modelo Predictivo G1.ipynb)

[Modelo Predictivo G1.ipynb - Colab]

D. Repositorio Proyecto:

Proyecto Modelo_Predictivo G1

XI. BIBLIOGRAFÍA:

Departamento Administrativo Nacional de Estadística (DANE). (2022). Encuesta Nacional de Calidad de Vida - ECV 2023. Recuperado de https://caldata.caldas.gov.co/wp-content/uploads/2023/05/Boletin_Calidad-de-Vida-2022-Marzo-2023.pdf

Organización de las Naciones Unidas (ONU). (2018). Declaración Universal de Derechos Humanos. Recuperado de https://news.un.org/es/story/2018/12/1447511

Universidad Externado de Colombia. (2019). Calidad de vida: un concepto más complejo de lo que parece. Recuperado de https://librepensador.uexternado.edu.co/calidad-de-vida-un-concepto-mas-complejo-de-lo-que-parece

Organización Mundial de la Salud (OMS). (2023). Calidad de vida, una meta difícil de alcanzar en Colombia. El Periódico UNAL. Recuperado de https://periodico.unal.edu.co/articulos/calidad-de-vida-una-meta-dificil-de-alcanzar-en-colombia

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2023). Factores determinantes de la calidad de vida. BBVA. Recuperado de https://www.bbva.com/es/salud-financiera/que-factores-determinan-la-calidad-de-vida-y-como-se-puede-mejorar

Departamento Nacional de Planeación - DNP. (2023). Pobreza Monetaria y Desigualdad. Dirección de Estudios Económicos. Recuperado de https://colaboracion.dnp.gov.co/CDT/PublishingImages/Planeacion-y-desarrollo/2024/Agosto/pdf/pobreza-monetaria.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia (MinTIC). (2023). Boletines de Conectividad. Recuperado de https://www.mintic.gov.co

Departamento Administrativo Nacional de Estadística (DANE). (2023). Encuesta de Calidad de Vida y Censo Nacional de Población y Vivienda. Recuperado de https://www.dane.gov.co

Banco Mundial. (2022). Informe sobre la brecha digital en América Latina. Recuperado de https://www.worldbank.org

Comisión Económica para América Latina y el Caribe (CEPAL). (2022). Transformación digital en América Latina y el Caribe: Agenda de políticas públicas. Recuperado de https://www.cepal.org/es

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2022). Digital Economy Outlook 2022. Recuperado de https://www.oecd.org

Naciones Unidas. (1948). Declaración Universal de Derechos Humanos. Recuperado de https://www.un.org/es/about-us/universal-declaration-of-human-rights

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2021). Informe de la UNESCO sobre Inclusión Digital y Educación. Recuperado de https://www.unesco.org

Programa de las Naciones Unidas para el Desarrollo (PNUD). (2022). Informe sobre el Desarrollo Humano 2022: La digitalización y el desarrollo humano. Recuperado de https://hdr.undp.org

Naciones Unidas. (1948). Declaración Universal de Derechos Humanos. Recuperado de https://www.un.org/es/about-us/universal-declaration-of-human-rights

La República. TECNOLOGÍA. Colombia se ubicó en el último lugar de países de la Ocde en cobertura de internet. (2023, mayo 21). Recuperado de https://www.larepublica.co/globoeconomia/colombia-se-ubico-en-el-ultimo-lugar-de-paises-de-la-ocde-en-cobertura-de-internet-3620379#:~:text=Seg%C3%BAn%20la%20Organizaci%C3%B3n%20para%20la,tiene%20acceso%20a%20este%20servicio.