PROYECTO FINAL BOOTCAMP IA

Tripulantes - Grupo No. 1
Maria Andrea 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.

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

Introducción

En Colombia la conectividad a internet es crucial para el desarrollo socioeconómico, sin embargo, persiste una marcada brecha digital entre zonas urbanas y rurales.

Esta investigación analiza cómo el acceso a internet influye en el desarrollo de la comunidades, a partir de un modelo de agrupamiento con base en los resultados de la Encuesta de Calidad de Vida -ECV 2023 - del DANE.

Imagen creada por ChatGPT

Planteamiento del problema

Este proyecto aborda cómo **la falta de acceso a internet limita el desarrollo socioeconómico** de las familias en Colombia, **afectando** factores como **la educación, los ingresos y la empleabilidad**. Utilizando un modelo de agrupación basado en inteligencia artificial, se busca identificar patrones entre conectividad y condiciones de vida para **orientar políticas de conectividad efectivas**. El objetivo es cerrar la brecha digital y promover un crecimiento inclusivo, mejorando la calidad de vida en distintas regiones del país.

Conectividad a internet y su impacto en el desarrollo socioeconómico

Acceso a internet

La disponibilidad y calidad de la conectividad a internet varía significativamente entre regiones de Colombia.

Oportunidades

Una mejor conectividad a internet abre puertas a nuevas oportunidades económicas y de educación.

Impacto

El estudio analiza cómo esta conectividad se relaciona con los indicadores de desarrollo de las familias.

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 "Encuesta de Calidad de Vida" del DANE 2023 para identificar las variables clave de conectividad y desarrollo socioeconómico que impactan la calidad de vida en distintas regiones de Colombia.

Describir las variables socioeconómicas y de conectividad, resaltando patrones que caracterizan a las familias según su nivel de acceso a internet y su situación socioeconómica.

Evaluar y aplicar modelos de agrupación como K-Means y jerárquicos para detectar patrones de desarrollo y conectividad en las comunidades.

Proponer recomendaciones para políticas de conectividad en las regiones con mayores necesidades, buscando reducir la brecha digital y fomenta el desarrollo inclusivo en todo el país.

Justificación

- Este proyecto busca identificar las desigualdades en el acceso a internet en Colombia, donde solo el 60,5% de la población tiene conexión. (La República, 2023)

- Las limitaciones de conectividad afectan especialmente a zonas rurales, con menos del 25% de acceso frente al 65% en áreas urbanas (DNP, 2023), lo que restringe oportunidades educativas, laborales y productivas, impactando directamente el desarrollo socioeconómico.

- Promover una Colombia más conectada y equitativa, a través de un modelo de agrupación que permita focalizar esfuerzos y acciones en educación, empleo y productividad y apoyando el desarrollo inclusivo y sostenible del país.

		Sin ayudas	Aporte Ayudas	Resultados 2023
Pobreza Monetaria	**Nacional**	37,4	-4,4	33,0
	Cabeceras	33,7	-3,1	30,6
	Centros poblados y rural disperso	49,5	-8,3	41,2
Pobreza Monetaria extrema	**Nacional**	16,1	-4,7	11,4
	Cabeceras	12,2	-3,3	8,9
	Centros poblados y rural disperso	29,2	-9,4	19,8

Fuente: DANE: Rueda de prensa pobreza monetaria.

Análisis de los factores clave que influyen en el desarrollo socioeconómico

Ingresos

Los ingresos familiares son un indicador fundamental del nivel de desarrollo.

Educación

El acceso y calidad de la educación impacta significativamente en el desarrollo.

Salud

Las condiciones de salud y acceso a servicios médicos son también factores clave.

Vivienda

El tipo y calidad de la vivienda reflejan el nivel socioeconómico de las familias.

Relación entre la conectividad a internet y el desarrollo socioeconómico

1

Acceso a Internet

Las regiones con mayor conectividad a internet tienden a tener un mayor desarrollo.

2

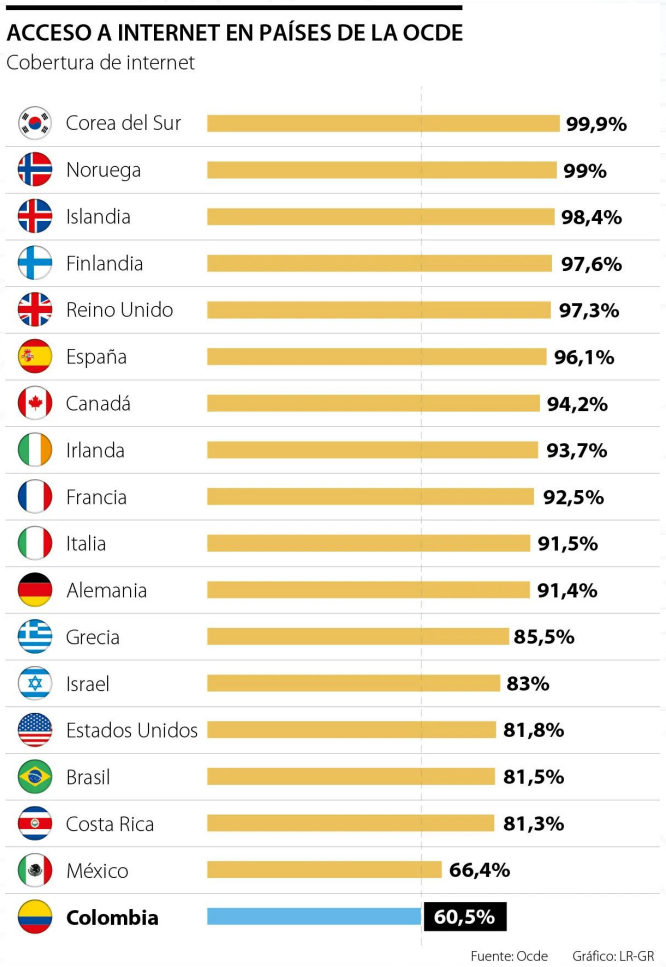
Oportunidades

La conectividad brinda acceso a educación, información y herramientas de productividad.

3

Desarrollo

Esto se traduce en mejores indicadores socioeconómicos para las familias conectadas.



- **Posición No.18**

Alcance

Bootcamp Talento Tech MinTIC, Enfoque desarrollo habilidades tecnológicas y promoción de la inclusión digital.

Contexto

Duración

Del 16 de septiembre hasta el 9 de noviembre 2024, en el marco del bootcamp

Limitación

No se desplegará en entornos reales ni se harán pruebas en campo.

Enfoque

Análisis y construcción del modelo sin intervención directa en comunidades.

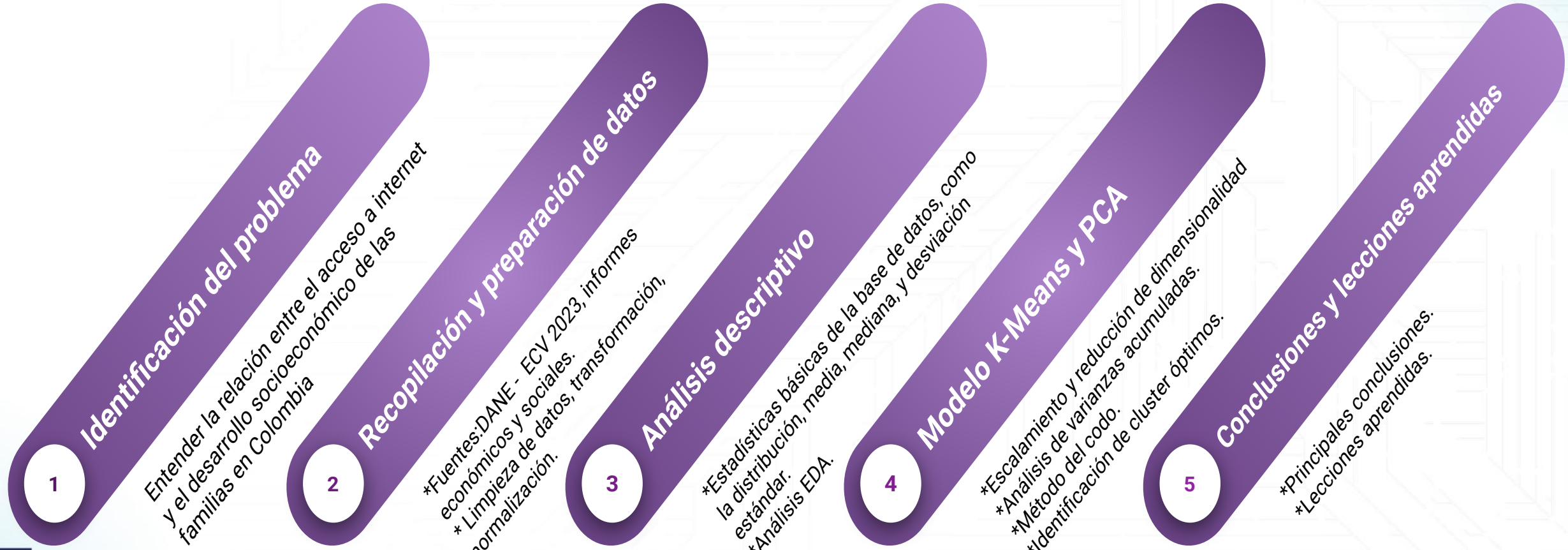
Uso IA

Solo para simulación y análisis de datos existentes.

Resultados

*Hallazgos y recomendaciones.
*Modelo y resultados para el bootcamp
y otras partes interesadas.

Metodología – Fases clave



Etapa - Recolección y preparación de datos

Características del Estudio ECV 2023

Características ECV 2023

*Caracterizar la población en los diferentes aspectos involucrados en el bienestar de los hogares.

*Acceso a bienes y servicios públicos, privados o comunales, salud, educación, atención integral de niños y niñas.

*Unidades de análisis : Hogares, familias e individuos de los municipios colombianos.

Etapas del estudio

*Unidades Primarias de Muestreo (UPM): Son todos los municipios del país.

*Unidades Secundarias de Muestreo (USM): Son conglomerados de 10 viviendas contiguas en promedio

Versión 1 (13 de abril de 2023)

Segmentos y variables

Segmentos:
*Educación : alfabetización, nivel educativo, asistencia escolar, subsidios.
*Conectividad a Internet : uso, frecuencia y puntos de acceso.
*Condiciones económicas : Ingresos del hogar, ocupación y percepción de pobreza.

Data: ECV 2023
*93 variables.
*223.695 Registros

Repositorio DANE - Catálogo Central de Datos

Estudios del DANE con respecto a calidad de vida y acceso a internet. -Encuesta de Calidad de Vida -
2021 -2023

Repositorio de datos oficial del DANE. www.microdatos.dane.go.co
https://microdatos.dane.gov.co/index.php/catalog/827/get-microdata

Etapa - Recolección y preparación de datos

Recolección

Se recopilaron datos a nivel nacional sobre indicadores socioeconómicos y de conectividad.

1

Limpieza

Los datos fueron depurados y estandarizados para el análisis posterior.

2

Modelado

Los algoritmos de IA y aprendizaje automático se aplicaron para generar conclusiones

3

```
#@title Verificar si hay datos faltantes
print("\nDatos faltantes por columna:")
print(bdCondicionesVida_df.isnull().sum().sum())
```

```
Datos faltantes por columna:
0
```

```
Tipos de datos por columna:
DIRECTORIO      int64
MPIO            int64
P6160           int64
P8586           int64
P8587           int64
                ...
P3516S6         int64
P3516S7         int64
P3516S8         int64
PROB_IAMG     float64
PROB_IAG      float64
Length: 93, dtype: object
```

Etapa - Análisis Exploratorio

Análisis de estadísticas descriptivas.

- Todas las variables tienen el mismo número de registros (223,695), lo cual indica que no hay datos faltante.
- La variable P8586 tiene un valor medio de 1.74, lo cual podría indicar una categorización binaria (1 y 2) con una inclinación hacia el valor 2. Las variables P1910 a P1912 con promedios cercanos a 5.
- Los valores de P8587 y P8587S1 alcanzan un máximo de 13, lo cual sugiere que estos datos pueden estar representando una escala de clasificación extensa o múltiples categorías dentro de la misma variable

	DIRECTORIO	MPIO	P6160	P8586	P8587	P8587S1	P1910	P1911	P1912	P1084
count	2.236950e+05	223695.000000	223695.000000	223695.000000	223695.000000	223695.000000	223695.000000	223695.000000	223695.000000	223695.000000
mean	8.087957e+07	47693.250368	1.080918	1.749695	3.697280	4.427761	4.537218	4.503172	4.908688	2.580903
std	2.026928e+07	29768.733955	0.272710	0.433190	2.735685	4.139877	1.163039	1.215432	0.544196	1.775312
min	7.910114e+07	5001.000000	1.000000	1.000000	1.000000	0.000000	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
50%	7.993241e+07	47245.000000	1.000000	2.000000	3.000000	4.000000	5.000000	5.000000	5.000000	2.000000
75%	8.162743e+07	73268.000000	1.000000	2.000000	5.000000	8.000000	5.000000	5.000000	5.000000	5.000000
max	8.201040e+08	99773.000000	2.000000	2.000000	13.000000	13.000000	5.000000	5.000000	5.000000	5.000000

Etapa - Análisis Exploratorio

Análisis de estadísticas descriptivas. (Algunas consideraciones)

P6160 y P1910

Homogeneidad en algunas variables

Sugiere que estos aspectos están bastante uniformes entre las observaciones.

P8587 y P8587S1

Alta variabilidad

Muestran alta variabilidad, lo cual es ideal para el análisis de clustering, ya que pueden aportar diferenciación entre los grupos o clusters

P1910, P1911, y P1912

Valores altos en variables de condiciones socioeconómicas

Sugiere una predominancia de condiciones en ciertos indicadores socioeconómicos en la muestra, lo que puede variar entre regiones.

Aspectos destacados del análisis exploratorio de datos (EDA)

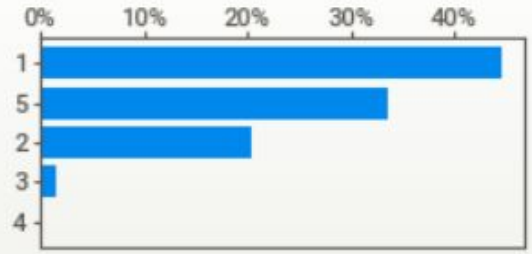
Hallazgos clave :

- Fuerte correlación entre variables socioeconómicas y acceso a Internet.
- Disparidades regionales, especialmente en la penetración de Internet en zonas rurales y urbanas.
- Mayor conectividad asociada a mayores niveles de educación y zonas urbanas.

P1084

VALORES: 223.695 (100%)
DESAPARECIDO: ---

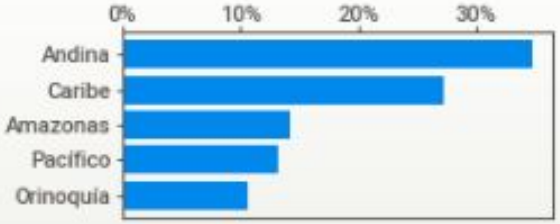
DISTINTO: 5 (<1%)



Region

VALUES: 223,209 (>99%)
MISSING: 486 (<1%)

DISTINCT: 5 (<1%)



Modelado de los datos

Aplicación de IA y Machine Learning para el análisis de datos

1 Algoritmos Avanzados

Se utilizaron técnicas de clustering, clasificación y predicción para analizar los datos.

2 Descubrimiento de Patrones

Los modelos de IA lograron identificar tendencias y relaciones ocultas en los datos.

3 Escalabilidad

El poder computacional de los algoritmos permitió analizar grandes volúmenes de información.

Modelado de los datos

Procesamiento de Datos

Escalamiento de los datos

El escalamiento ayuda a ajustar las variables a un rango similar, generalmente entre 0 y 1 o con una media de 0 y distorsión estándar de 1.

Partición de los datos

La partición se realizó con base en:
70% para entrenamiento
30% para prueba

Binary Encoding

Este tipo de codificación transforma variables categóricas (como ***nombres de ciudades o departamentos***) en una representación binaria que es más adecuada para ciertos algoritmos de aprendizaje automático, especialmente cuando hay muchas categorías.

Varianza Acumulada

*La varianza acumulada es útil para reducir la dimensionalidad de los datos y optimizar el modelo de agrupación.

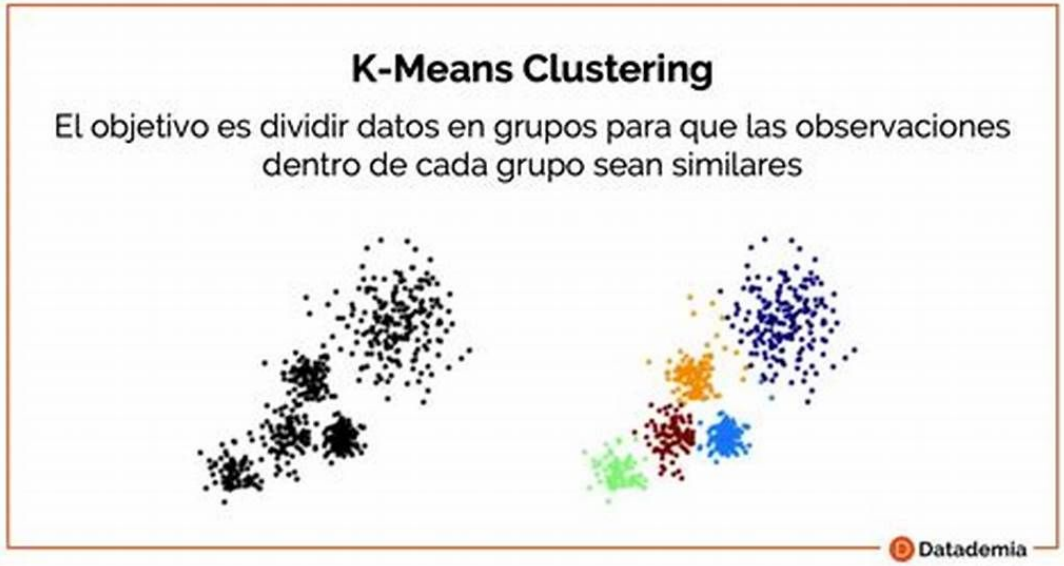
*Para el modelo se tomaron los datos que explican hasta el 80% de la varianza acumulada.

*Este proceso optimiza los datos para el modelo de agrupación.

Modelado de los datos

Modelo de agrupamiento: K-means

- **Objetivo** : Agrupar regiones por similitud socioeconómica y nivel de conectividad.
- **Algoritmo** : K-Means y agrupamiento jerárquico.
- **Métricas clave** : Uso del Internet, Acceso a internet, Alfabetismo, ingresos, desempleo, tasa de pobreza.



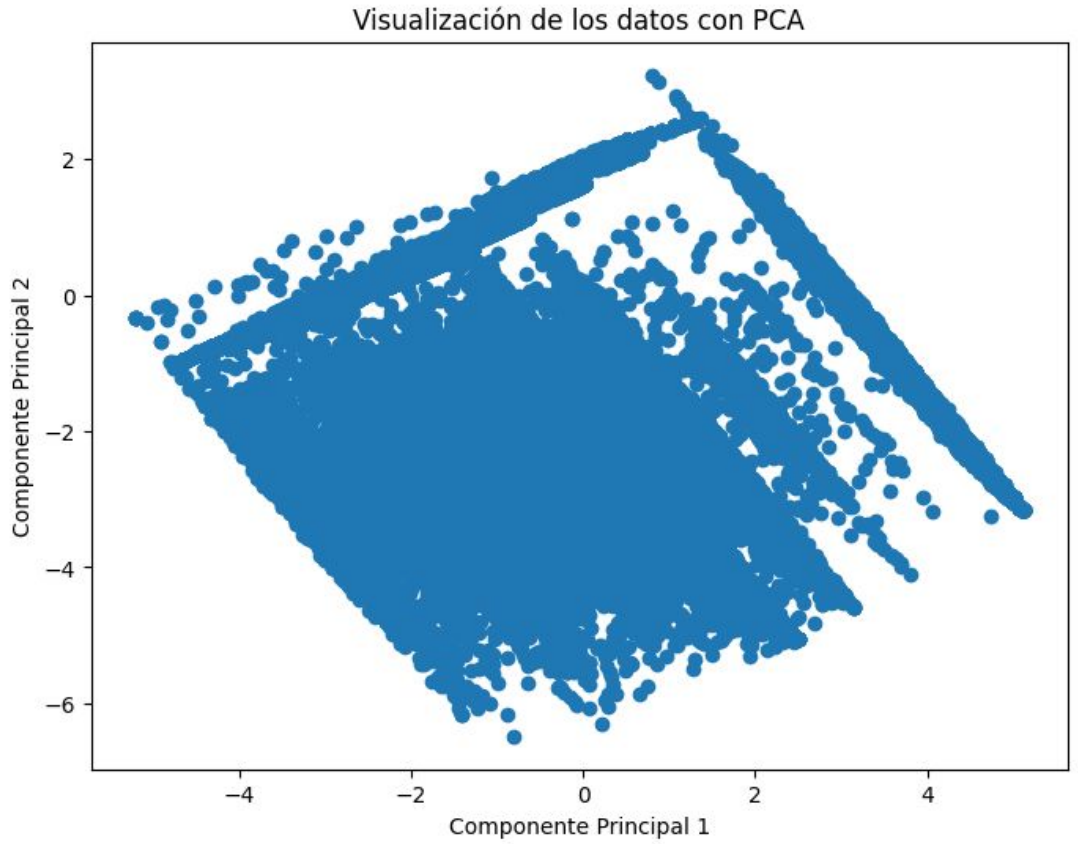
Perspectivas clave y desigualdades regionales

- Se identificaron regiones con acceso limitado a Internet y menor desarrollo socioeconómico.
- Se destacaron áreas de impacto potenciales para las políticas públicas para mejorar la conectividad a Internet y reducir la desigualdad.

Modelado de los datos

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, especialmente en conjuntos de datos con muchas características.
- A parti dos componentes principales (n_components = 2), se representa el conjunto de datos en dos dimensiones, capturando la mayor cantidad de información posible con menos variables.
- En conjunto, estos dos componentes explican alrededor del 60.2% de la varianza total. Esto significa que más de la mitad de la información original está conservada en los datos reducidos.



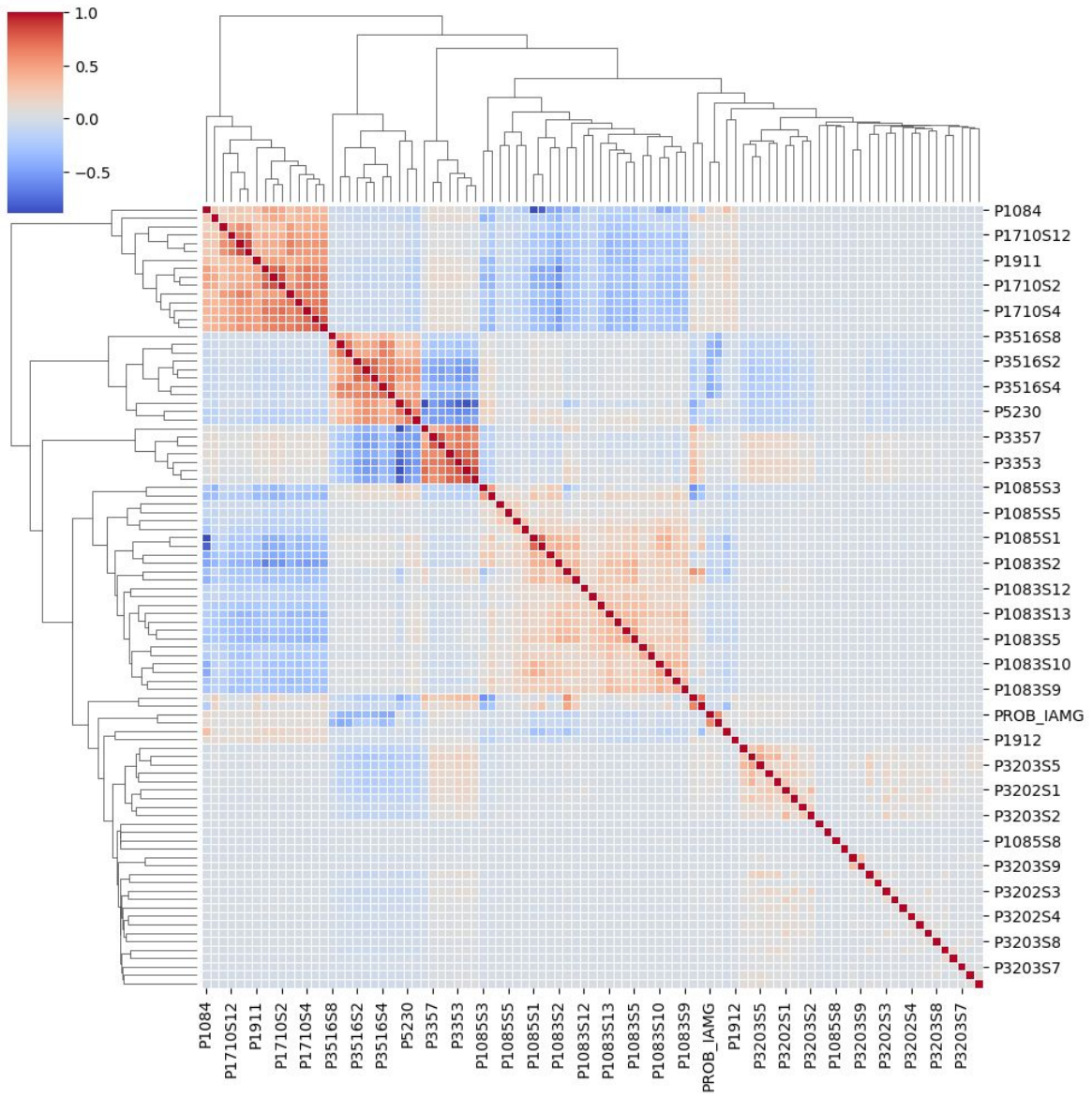
Modelado de los datos

Representación de clúster jerárquico

*Para la gráfica los tonos rojos indican una correlación positiva (es decir, cuando una variable aumenta, la otra también tiende a aumentar), mientras que los tonos azules muestran una correlación negativa (cuando una variable aumenta, la otra tiende a disminuir).

*Las agrupaciones formadas sugieren conjuntos de variables que tienden a comportarse de manera similar, lo que puede revelar patrones relevantes para el análisis del impacto de la conectividad.

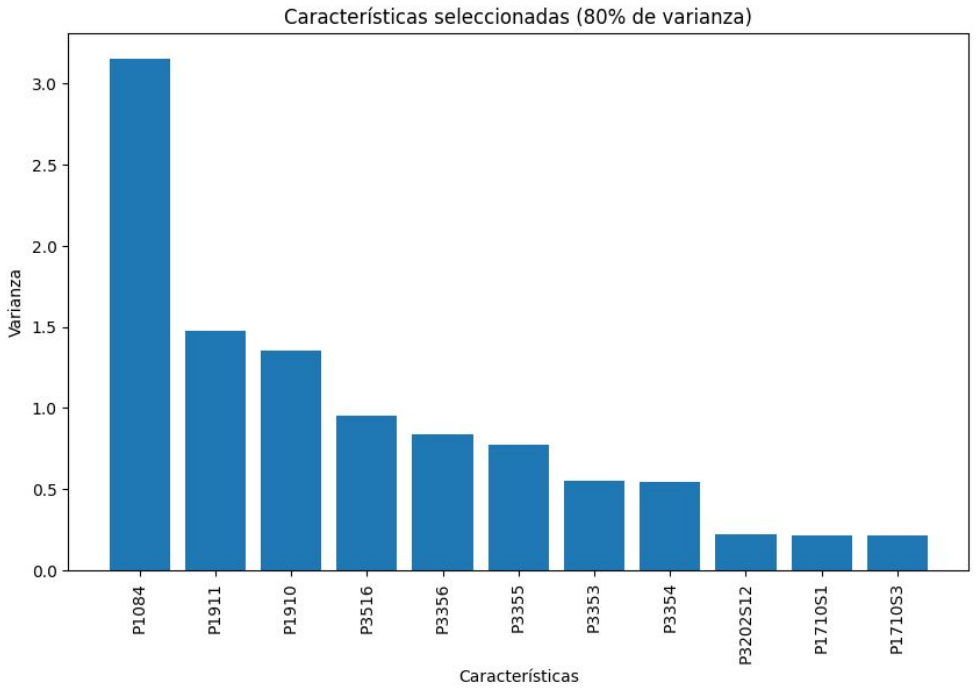
*Lo anterior genera la necesidad de realizar un proceso que mejore la dimensionalidad de las variables a partir de la funcionalidad ***category_encoders***



Resultados

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.

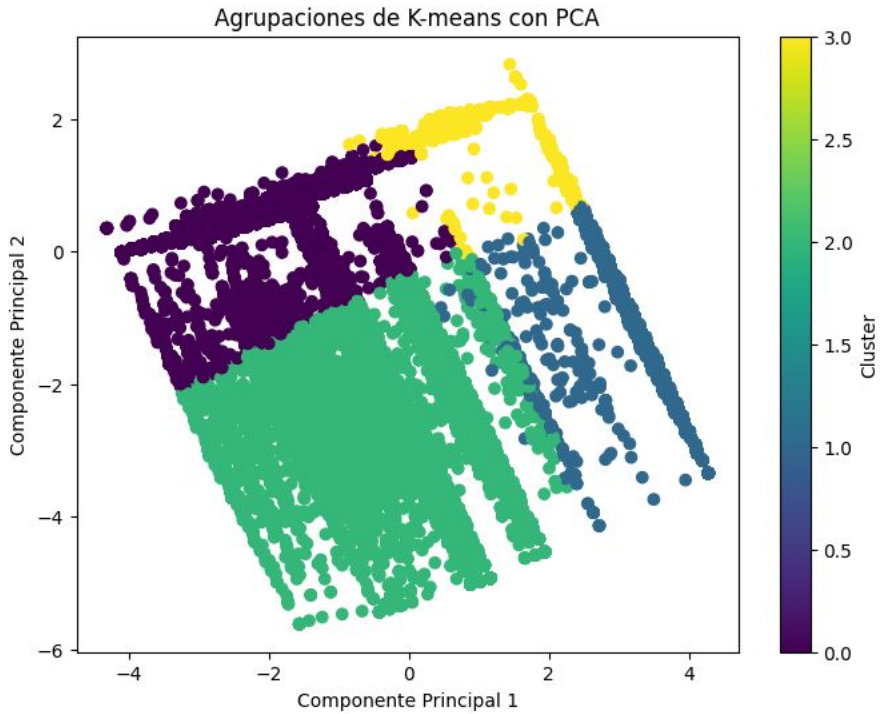
Al reducir las variables a solo aquellas que explican el 80% de la varianza, el modelo se vuelve más interpretativo, lo que facilita la identificación de patrones de conectividad y desarrollo socioeconómico en los distintos grupos de familias.

Resultados

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.

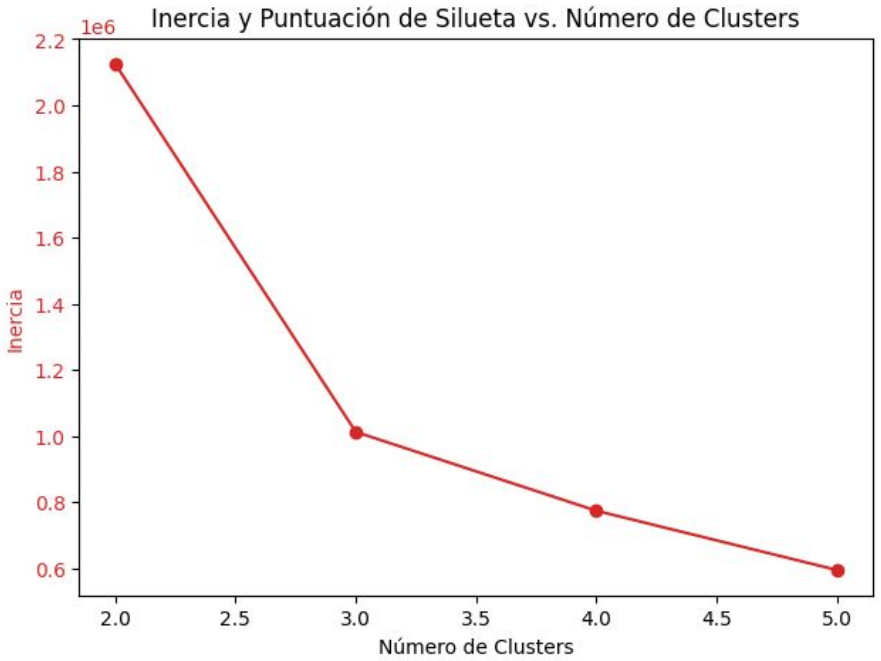


a. **Cluster de alta frecuencia de uso de Internet.** Los hogares en este cluster presentan un uso de Internet frecuente, lo que sugiere que la conectividad es una parte fundamental de su vida cotidiana.
b. **Cluster de uso intermitente o moderado de Internet.** Este cluster representa hogares que tienen acceso a Internet, pero su uso es intermitente o moderado.
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.

Resultados

Gráfica de inercia Vs Número de Cluster.

La gráfica muestra 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:



Número Óptimo de Clusters: La gráfica sugiere que **3 clusters** es un buen equilibrio, ya que es el punto donde la disminución de inercia empieza a ser menos pronunciada (punto de "codo").

Reducción de Inercia: La inercia disminuye significativamente al pasar de 2 a 3 clusters, pero después de este punto, la reducción es más gradual, lo que indica menores ganancias en compacidad al agregar más clusters.

Simplicidad vs. Precisión: Elegir 3 clusters proporciona una buena relación entre compactación de los datos y simplicidad del modelo, evitando un aumento innecesario de la complejidad al agregar más clusters.

Conclusiones

- 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.

- 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.

- El uso de técnicas de reducción de dimensionalidad y selección de características permitió que el modelo se centrará en las variables más significativas, mejorando la interpretación de los resultados.

Conclusiones

- 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.

- 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.

Conclusiones

1 Brecha Digital

Aún existe una brecha digital significativa, que afecta el desarrollo de muchas familias.

2 Oportunidades

Mejorar la conectividad a internet puede abrir nuevas puertas de oportunidad en la educación en la empleabilidad y la productividad.

3 Políticas

Se requieren políticas e inversiones para cerrar la brecha y fomentar el desarrollo.

El estudio demuestra la importancia de la conectividad a internet como facilitadora del desarrollo socioeconómico en Colombia. Estos hallazgos pueden informar políticas y estrategias para promover la inclusión digital y mejorar el bienestar de las familias.

Lecciones aprendidas

- La IA tiene un amplio campo de desarrollo en los diferentes ámbitos del conocimiento, lo que representa una valiosa oportunidad como nueva ventaja competitiva en la era digital para las organizaciones.
- A través de modelos de IA es posible atender necesidades sociales de manera efectiva con menores recursos y mayor productividad desde los resultados.
- El desarrollo de los modelos de IA aplicado a las diferentes industrias en el país presenta un desarrollo básico con respecto al potencial en la resolución de problemas a gran escala.
- Fortalecer las habilidades técnicas específicas requeridas para el desarrollo del proyecto, como programación, análisis de datos, o el uso de herramientas especializadas como Chat GPT, representan un gran logro para el equipo de trabajo.
- Contar con los conceptos generales entorno a los modelos de Machine Learning, uso y aplicación, permitirán en el corto y mediano plazo facilitar nuevas iniciativas desde la particularidad de los roles de cada miembro en el equipo de trabajo.

Anexos

COSTOS DEL PROYECTO	
ITEM	**VALOR**
Papelería	$200.000
Internet	$258.000
Servicio públicos	$360.000
Asesoría Expertos Técnicos	$9.600.000
Procesamiento y Análisis de datos	$12.000.000
TOTAL	**$22.418.000**

Anexos

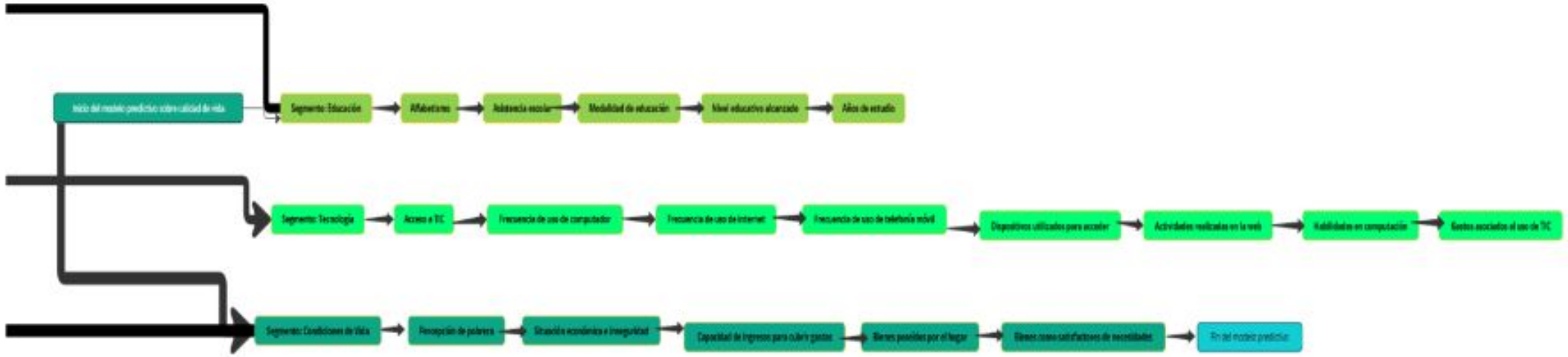
- **Cuaderno en Colab (Modelo Predictivo G1.ipynb)**

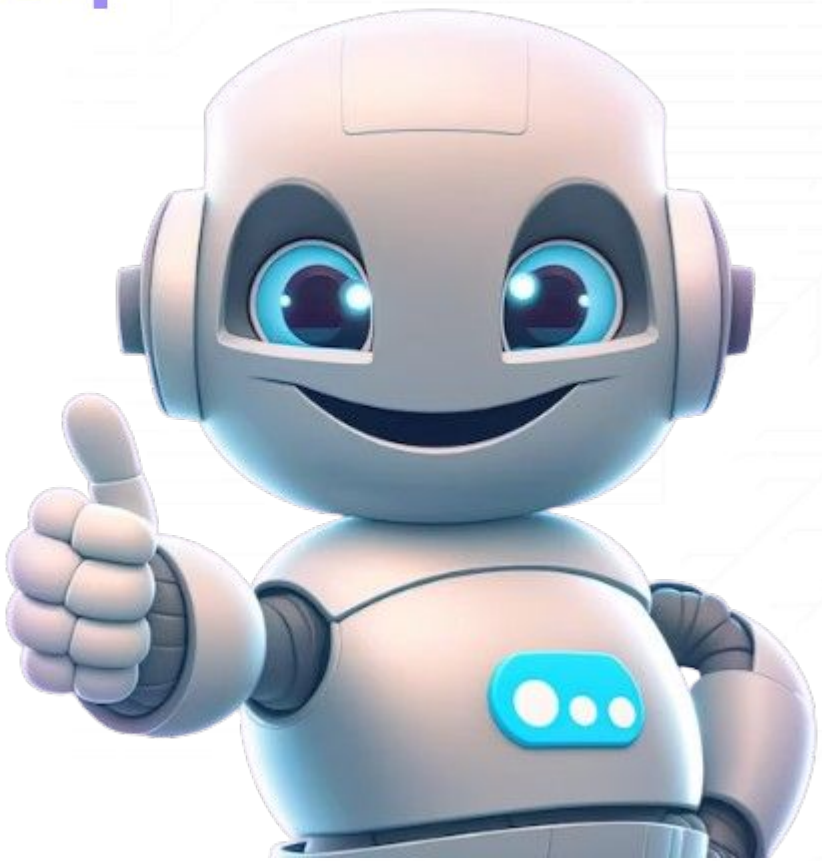
  Modelo Predictivo G1.ipynb - Colab

- **Repositorio Proyecto:**

  Proyecto Modelo_Predictivo G1

- **Diagrama de flujo del diccionario de datos:**





Gracias