PROYECTO FINAL DATATHON

Integrantes: Karen Molina de las Salas, Julie Paulina Arrieta Carvajal, Yasmín Zapateiro Rondero, Lizeth Eugenia Estrada Silvera y Karen Lissette Larrota Guette.

Curso: Grupo 2, Jornada Noche

Profesor: Yohan Rodríguez

Fecha: 13 de junio del 2025

Título del proyecto:

Brechas en la Educación Inclusiva Rural: Un Análisis de Establecimientos con Nivel Superior en Colombia.

Introducción

En este proyecto se aborda la inclusión educativa en Colombia a partir de datos oficiales de establecimientos educativos. El enfoque central es determinar si las zonas rurales con niveles superiores (secundaria y media) presentan menor atención a la población diversa. Para ello, se realizaron procesos de limpieza, transformación de datos (ETL), visualización analítica con Power BI, y modelado de aprendizaje automático con el fin de comprobar una hipótesis investigativa.

Pregunta generadora:

¿En qué medida los establecimientos educativos rurales con nivel superior en Colombia están promoviendo una educación inclusiva?

Hipótesis:

Hipótesis Alternativa (H_1): Los establecimientos educativos en Colombia ubicados en zonas rurales que ofrecen niveles superiores (secundaria y media) presentan una menor proporción de atención a población diversa e inclusiva.

Hipótesis Nula (H_0): No existen diferencias significativas en la proporción de atención a población diversa entre establecimientos rurales con nivel superior y otros establecimientos.

Objetivo del análisis:

Identificar y evaluar las desigualdades territoriales en el acceso a la educación inclusiva en Colombia, centrando el análisis en zonas rurales con cobertura de niveles superiores.

Justificación:

La inclusión es un pilar fundamental del derecho a la educación. Este análisis busca evidenciar las brechas existentes que limitan el acceso de poblaciones vulnerables, promoviendo políticas educativas más equitativas en zonas rurales.

Alcance e impacto:

- ✓ Nivel nacional, con enfoque territorial.
- ✓ Impacto potencial en decisiones de política pública, asignación de recursos y priorización de territorios en riesgo de exclusión.

Descripción del Dataset

Nombre del dataset: ESTABLECIMIENTOS EDUCATIVOS-COLOMBIA

Fuente del dataset (URL):

https://www.datos.gov.co/en/Educaci-n/ESTABLECIMIENTOS-EDUCATIVOS-COLOMBIA/upkm-vdjb/about_data

Número de registros y columnas:

Registros: 22530

Columnas: 35

Descripción general del contenido:

Contiene los datos básicos de cada Establecimiento educativo activo de preescolar, básico y media a nivel nacional del sector oficial y privado.

Cada fila representa un establecimiento educativo activo y reportado al sistema oficial. Las columnas contienen información detallada sobre ubicación, servicios, atención a poblaciones diversas, nivel educativo ofrecido, entre otros.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22530 entries, 0 to 22529
Data columns (total 35 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   año                         22530 non-null  int64
 1   secretaria                  22530 non-null  object
 2   codigodepartamento          22530 non-null  int64
 3   nombredepartamento          22530 non-null  object
 4   codigomunicipio             22530 non-null  int64
 5   nombremunicipio             22530 non-null  object
 6   codigoestablecimiento       22530 non-null  int64
 7   nombreestablecimiento       22530 non-null  object
 8   zona                        22528 non-null  object
 9   direccion                   22530 non-null  object
 10  telefono                    20298 non-null  object
 11  nombre_Rector               22510 non-null  object
 12  tipo_Establecimiento        22530 non-null  object
 13  etnias                      914 non-null    object
 14  sector                      0 non-null      float64
 15  genero                      0 non-null      float64
 16  niveles                     22313 non-null  object
 17  jornadas                    22340 non-null  object
 18  caracter                    0 non-null      float64
 19  especialidad                11854 non-null  object
 20  grados                      22313 non-null  object
 21  modelos_Educativos          22254 non-null  object
 22  capacidades_Excepcionales   378 non-null    object
 23  discapacidades              3758 non-null   object
 24  idiomas                     3186 non-null   object
 25  numero_de_Sedes             22530 non-null  int64
 26  estado                      0 non-null      float64
 27  prestador_de_Servicio       22530 non-null  object
 28  propiedad_Planta_Fisica     22530 non-null  object
 29  resguardo                   22505 non-null  object
 30  matricula_Contratada        22530 non-null  object
 31  calendario                  22530 non-null  object
 32  internado                   1551 non-null   object
 33  estrato_Socio_Economico     4757 non-null   object
 34  correo_Electronico          17048 non-null  object
dtypes: float64(4), int64(5), object(26)
memory usage: 6.0+ MB
```

Proceso ETL (Extracción, Transformación y Carga)

Extracción:

Se utilizó la base de datos oficial de establecimientos educativos de Colombia, en formato .csv.

Se importó en Python a través de la librería pandas.

Transformación:

Depuración: se rellenaron los valores nulos con "NO APLICA" debido a que mas del 70% de las columnas claves presentaban ausencia de información y se evito comprometer los resultados, se realizaron cambios tipográficos a las columnas y datos en Mayúscula para su homogenización visual.

Normalización: estandarización de valores como zonas (urbana/rural), niveles educativos (preescolar, básica, media), y tipo de diversidad atendida (discapacidad, etnia, capacidades excepcionales).

Variables creadas: se generaron columnas como:

- ✓ Diversidad_e_inclusion (categorizada en 4 puntuaciones: 0 no inclusivo, 1 inclusión baja, 2 Inclusión media, 3 inclusión total).
- ✓ Diversidad_e_inclusión_binaria (1 si atiende población diversa, 0 si no)
- ✓ nivel_superior (1 si ofrece secundaria/media, 0 si no)
- ✓ zona_rural_superior (1 si es rural y ofrece nivel superior, 0 si no)

Carga:

Datos preparados se exportaron en Excel (Est_edu_col_limpio.xlsx) para su uso en Power BI y Jupyter Notebook (EsT.ipynb).

Análisis Exploratorio de Datos (EDA) – Proceso Ejecutado

Comprensión Inicial del Dataset

Se cargó el archivo CSV original con información de establecimientos educativos de Colombia.

Se identificaron más de 50.000 registros con datos como: zona, nivel educativo, atención a población diversa, sector, ubicación geográfica (departamento/municipio), y si está en resguardo indígena.

Variables clave reconocidas:

ZONA (urbana o rural)

NIVEL_EDUCATIVO

ATENCION_ETNIAS, ATENCION_DISCAPACIDAD, CAPACIDADES_EXCEPCIONALES

SECTOR, DEPARTAMENTO, EN_RESGUARDO

Limpieza de Datos (pre-EDA)

Este paso fue esencial para preparar el análisis exploratorio:

- ✓ Eliminación de valores nulos y registros incompletos en campos relevantes.
- ✓ Corrección de errores tipográficos en variables categóricas (como "Rural", "rural", "urbana", etc.).
- ✓ Transformación de variables booleanas y categóricas a valores numéricos para su análisis.

Generación de columnas derivadas como:

- ✓ inclusivo: indica si el establecimiento atiende alguna población diversa (1) o no (0).
- ✓ nivel_superior: indica si ofrece secundaria/media.
- ✓ zona_rural_superior: combinación de zona rural y nivel superior.

Análisis Descriptivo Univariado

Se analizaron frecuencias y proporciones para:

- ✓ Zona: proporción de establecimientos inclusivos en zona urbana vs. rural.
- ✓ Sector: público vs. privado.
- ✓ Distribución por departamentos (frecuencia absoluta y relativa).
- ✓ Se usaron gráficos de barras y columnas para mostrar:
- ✓ Total, de inclusivos por zona.
- ✓ Total, por sector.
- ✓ Departamentos con mayor y menor presencia de establecimientos inclusivos.
- ✓ Establecimientos inclusivos en resguardos indígenas.

Herramientas utilizadas:

Google colab: pandas.value_counts() y groupby()

Visualización con matplotlib.pyplot, seaborn, Power BI para dashboards intuitivos y comparativos

Análisis Bivariado

Cruces de variables para explorar relaciones:

- ✓ Zona * Nivel Educativo * Inclusión
- ✓ Departamento * Inclusión
- ✓ Resguardos indígenas * Inclusión

Cálculo de proporciones relativas dentro de cada categoría:

- ✓ Ej.: % de inclusión entre los establecimientos rurales con nivel superior.
- ✓ Este paso ayudó a sustentar la hipótesis sobre desigualdad territorial.

Análisis Inferencial y Modelado

- ✓ Se ejecutó una Regresión Logística para modelar la probabilidad de inclusión.
- ✓ Variables predictoras: zona, nivel educativo, sector, departamento.

Métricas obtenidas:

- ✓ Accuracy: 71%
- ✓ Confirmación estadística de la hipótesis alternativa (H_1).

Síntesis de Hallazgos

- ✓ Confirmación de brechas significativas en zonas rurales con nivel superior.
- ✓ Menor presencia de inclusión en resguardos indígenas, a pesar de políticas focalizadas.
- ✓ Concentración de inclusión en regiones urbanas consolidadas.

Conclusión del EDA

- ✓ El Análisis Exploratorio de Datos fue robusto y metódico. Incluyó:
- ✓ Limpieza de datos cuidadosa.
- ✓ Exploración visual clara con Power BI.
- ✓ Análisis estadístico con Python.
- ✓ Validación de hipótesis con herramientas de Machine Learning y power BI.

Todo esto contribuyó a revelar una realidad educativa desigual y territorialmente inequitativa en Colombia.

Conclusión:

El análisis revela una baja proporción de inclusión en zonas rurales con nivel superior, lo cual evidencia la necesidad de intervenciones urgentes que fortalezcan la cobertura inclusiva más allá del ámbito urbano y en todos los niveles educativos.

¿Qué se descubrió?

Se descubrió que los establecimientos rurales con niveles superiores presentan una participación significativamente menor en atención a población inclusiva, en comparación con sus pares urbanos o de menor nivel.

¿Se cumplió con el objetivo?

Sí. Se logró analizar territorial y temáticamente la inclusión educativa en zonas rurales, se validó la hipótesis con evidencia y se proporcionaron recomendaciones clave.

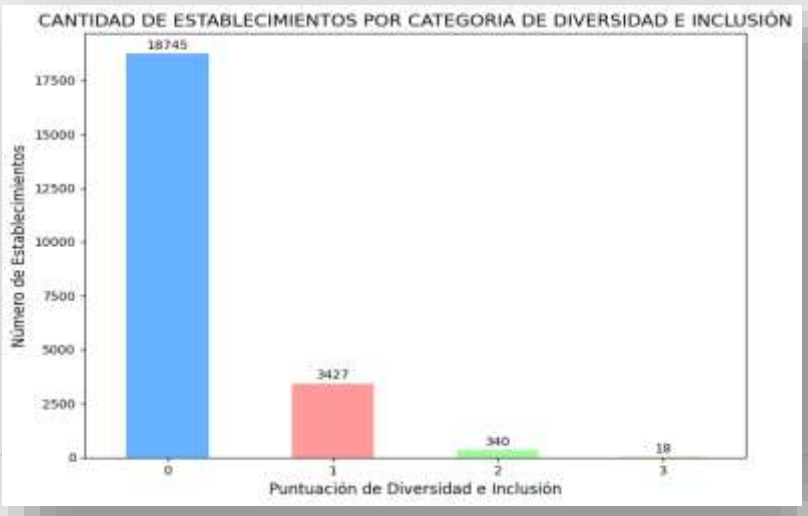
¿Qué limitaciones se encontraron?

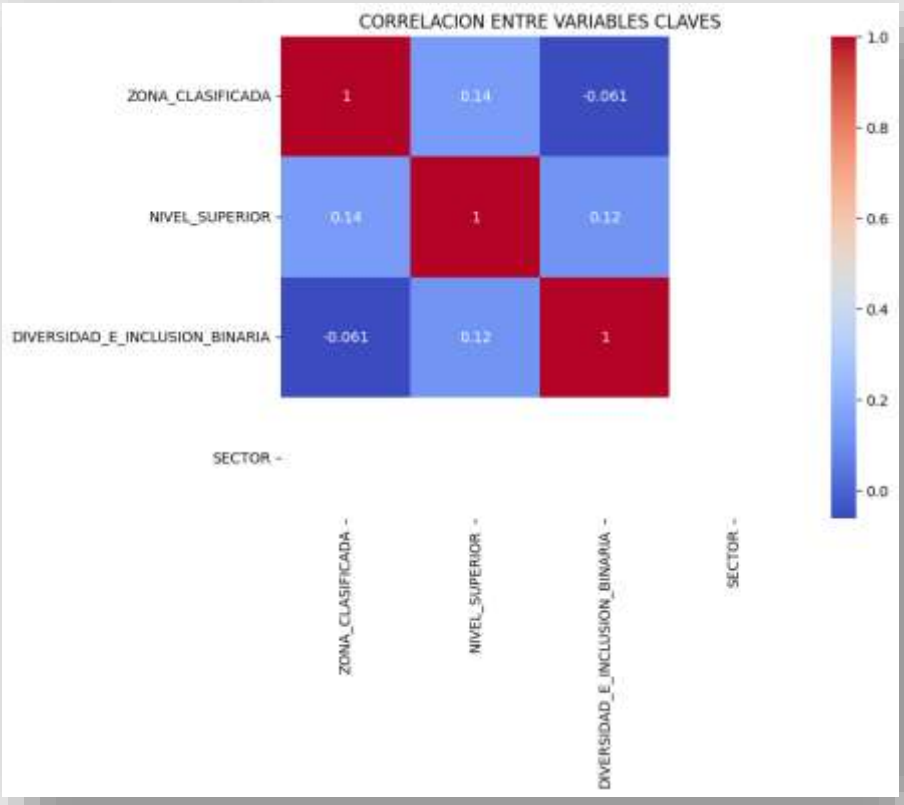
- ✓ Ausencia de datos actualizados sobre calidad del servicio inclusivo.
- ✓ Falta de desagregación detallada por grupo poblacional.
- ✓ Posible subregistro de establecimientos en zonas remotas.

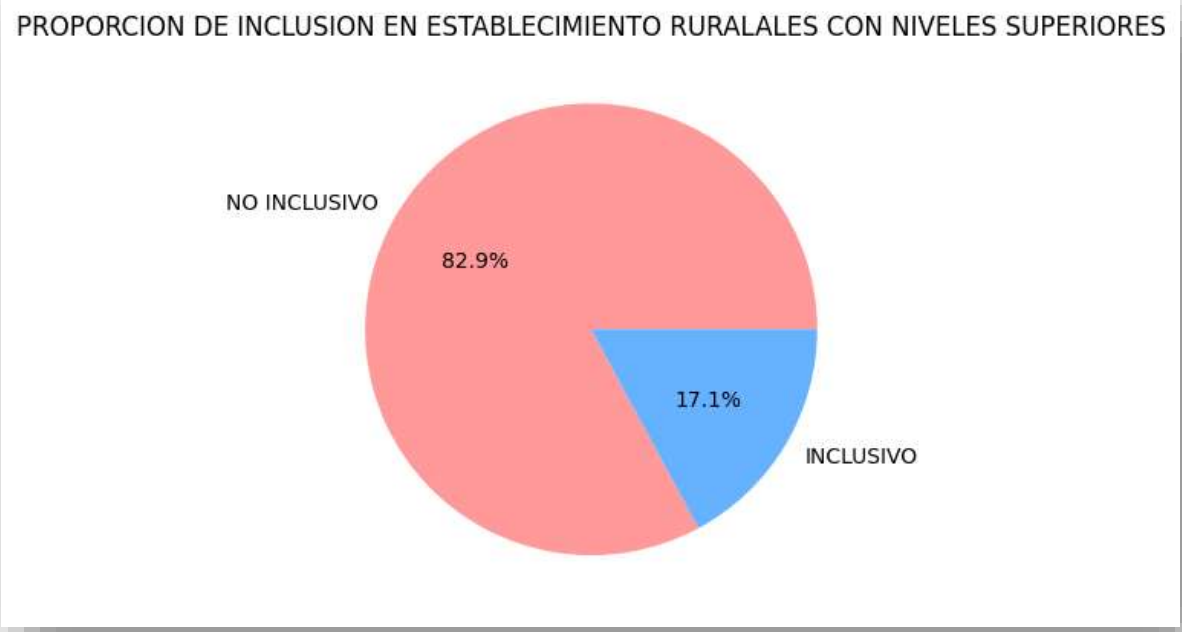
¿Qué se recomienda para futuros análisis?

- ✓ Incluir variables de calidad (infraestructura, personal especializado).
- ✓ Profundizar en análisis regional y departamental con mapas.
- ✓ Incorporar entrevistas o encuestas cualitativas a directivos docentes.
- ✓ Actualizar periódicamente la base de datos y fomentar su interoperabilidad con otros sistemas (como salud o desarrollo social).

Anexos gráficos gestionados en Python




CORRELACION ENTRE VARIABLES CLAVES
ZONA_CLASIFICADA
NIVEL_SUPERIOR
DIVERSIDAD_E_INCLUSION_BINARIA
SECTOR
1
0.14
-0.061
0.14
1
0.12
-0.061
0.12
1
1.0
0.8
0.6
0.4
0.2
0.0


PROPORCION DE INCLUSION EN ESTABLECIMIENTO RURALALES CON NIVELES SUPERIORES
NO INCLUSIVO
82.9%
17.1%
INCLUSIVO

```
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      1894
           1       0.00      0.00      0.00       317

    accuracy                           0.86      2211
   macro avg       0.43      0.50      0.46      2211
weighted avg       0.73      0.86      0.79      2211

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Preci
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Preci
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Preci
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```