



LOS OROYANOS

# CENTROS DE EDUCACIÓN MAL UBICADOS

Data mining

Defensoría del Pueblo identifica deficiencias en  
colegios públicos y privados de todo el distrito  
de Río Negro

11:24 AM 22/06/2023



Defensoría del Pueblo: gobiernos regionales y locales deben cumplir normas técnicas sobre ubicación de colegios

9:58 AM 25/04/2023



11:48 AM 29/03/2023



- El 42 % de instituciones educativas no cumplen con las normas de construcción y equipamiento vigentes.
- El 97.65 % de establecimientos de enseñanza no se encuentran de acuerdo con las normas de construcción y equipamiento vigentes.



- Aproximadamente, 1 de cada 10 escuelas pú

Defensoría del Pueblo advierte que 28 colegios públicos se encuentran en alto riesgo en el Callao

11:00 AM 01/03/2023



- Durante supervisión se comprobó cercos perimetéricos colapsados, columnas rajadas, paredes con salitre, así como módulos educativos antiguos que requieren ser sustituidos.
- Miles de estudiantes se verán afectadas/os de regresar a sus aulas en mal estado.

# PREOCUPANTE

Los **colegios mal ubicados**, situados en zonas remotas y carentes de servicios básicos, **impactan negativamente** en el rendimiento y bienestar de los estudiantes, aumentando el **riesgo de abandono y problemas sociales**.

Un estudio de la **UNESCO (2019)** reveló que estos colegios carecen de recursos educativos, limitando el acceso a una educación de calidad. **Mejorar la ubicación y calidad de estos centros es esencial para garantizar el derecho a la educación de todos los niños.**



# Compreensión del negocio

## OBJETIVO

Mejorar la educación a través de una mejor ubicación de centros educativos.

## STAKEHOLDERS



PERÚ

Ministerio  
de Educación

## PREGUNTA CLAVE

¿Es posible identificar ubicaciones óptimas para los centros educativos en Perú aplicando algoritmos de geolocalización y sistemas de información geográfica, de modo que se mejore la distribución de la oferta educativa?



# Comprendión de los datos

MINEDU

Dataset con los datos de las instituciones  
educativas en el Perú al 2022



Ministerio  
de Educación

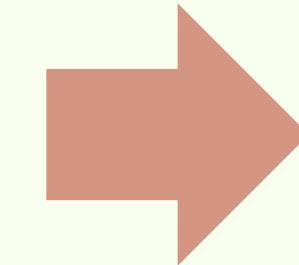


**OPEN BUILDINGS (GOOGLE RESEARCH)**

Dataset de huellas de edificios de África, Asia,  
América Latina y el Caribe,

# Dataset Minedu

Instituciones y programas  
educativos en el Perú.



- Descripciones de los niveles y modalidades educativas (inicial, primaria, secundaria, etc)
- Información sobre el tipo de gestión de las instituciones (pública, privada, etc).
- Incluye datos de contacto detallados de cada institución/local educativo.
- Cuenta con códigos geográficos y coordenadas que permiten identificar la localización exacta.
- Tiene atributos que indican el departamento, provincia y distrito de cada servicio educativo.

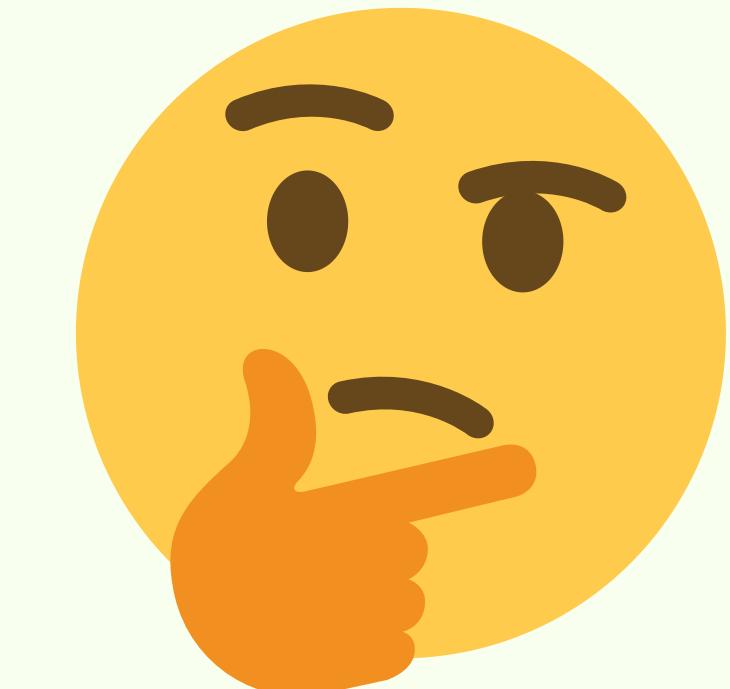
# Dataset Minedu

173,708 registros  
51 atributos

	Columna	Valores Únicos	Total	Valores NaN	Tipo de Dato
40	fecha_act	1	166522	0	object
38	d_estado	2	166522	0	object
37	estado	2	166522	0	int64
23	dareacenso	2	166522	0	object
22	area_censo	2	166522	0	int64
5	d_forma	3	166522	0	object
6	tipssexo	3	166522	0	int64
7	d_tipssexo	3	166522	0	object
8	gestion	3	166522	0	int64
9	d_gestion	3	166522	0	object
36	d_cod_tur	7	166522	0	object
35	cod_tur	7	166522	0	int64
0	anexo	8	166522	0	int64
10	ges_dep	10	166522	0	object
11	d_ges_dep	10	166522	0	object
33	tipoprog	15	47898	118624	float64
34	d_tipoprog	15	47898	118624	object
4	d_niv_mod	19	166522	0	object
25	d_dpto	25	166522	0	object
28	d_region	26	166522	0	object
3	niv_mod	27	166522	0	object
26	d_prov	196	166522	0	object
30	d_dreugel	238	166522	0	object

29	codoii	239	166522	0	int64
15	pagweb	946	1785	164737	object
27	d_dist	1722	166522	0	object
24	codgeo	1874	166522	0	int64
14	email	7414	10256	156266	object
39	fechareg	13502	144043	22479	object
18	localidad	21006	80567	85955	object
13	telefono	22048	36237	130285	object
17	referencia	26557	39655	126867	object
21	cen_pob	29238	166522	0	object
19	codcp_inei	31143	149113	17409	float64
20	codccpp	37093	165172	1350	float64
2	cen_edu	77851	166522	0	object
12	director	81426	108355	58167	object
16	dir_cen	81440	166522	0	object
1	codlocal	87180	118624	47898	float64
32	nlong_ie	87409	166522	0	float64
31	nlat_ie	88669	166522	0	float64

Toda esa  
información es  
relevante?



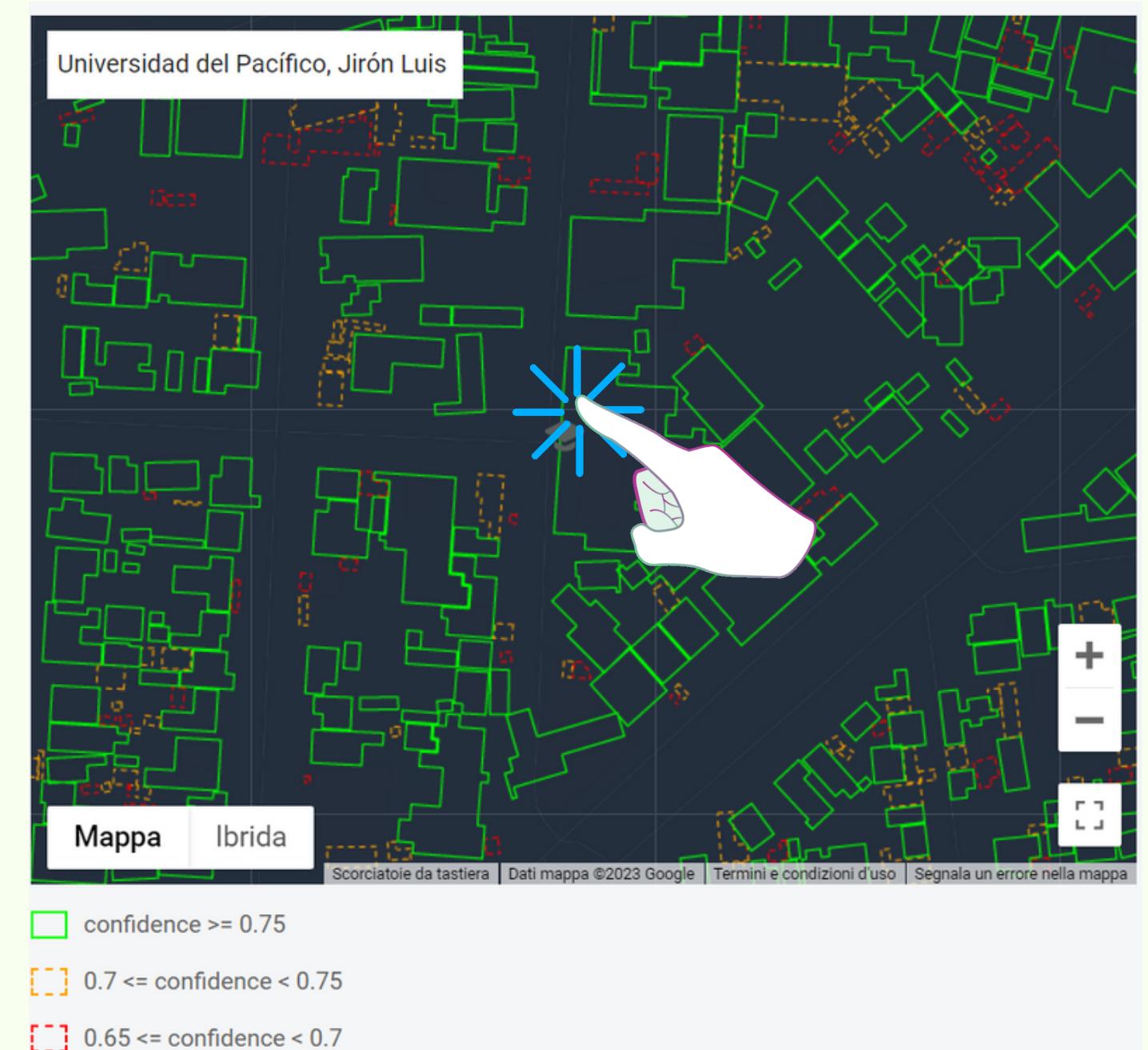
# ¿Qué es open buildings?

Conjunto de **datos open source** de huellas de edificios derivadas de **imágenes satelitales**.

Fue creado como parte del proyecto Open Buildings por Google Research para apoyar aplicaciones de bien social.

Contiene 1.8 billones de detecciones de edificios.

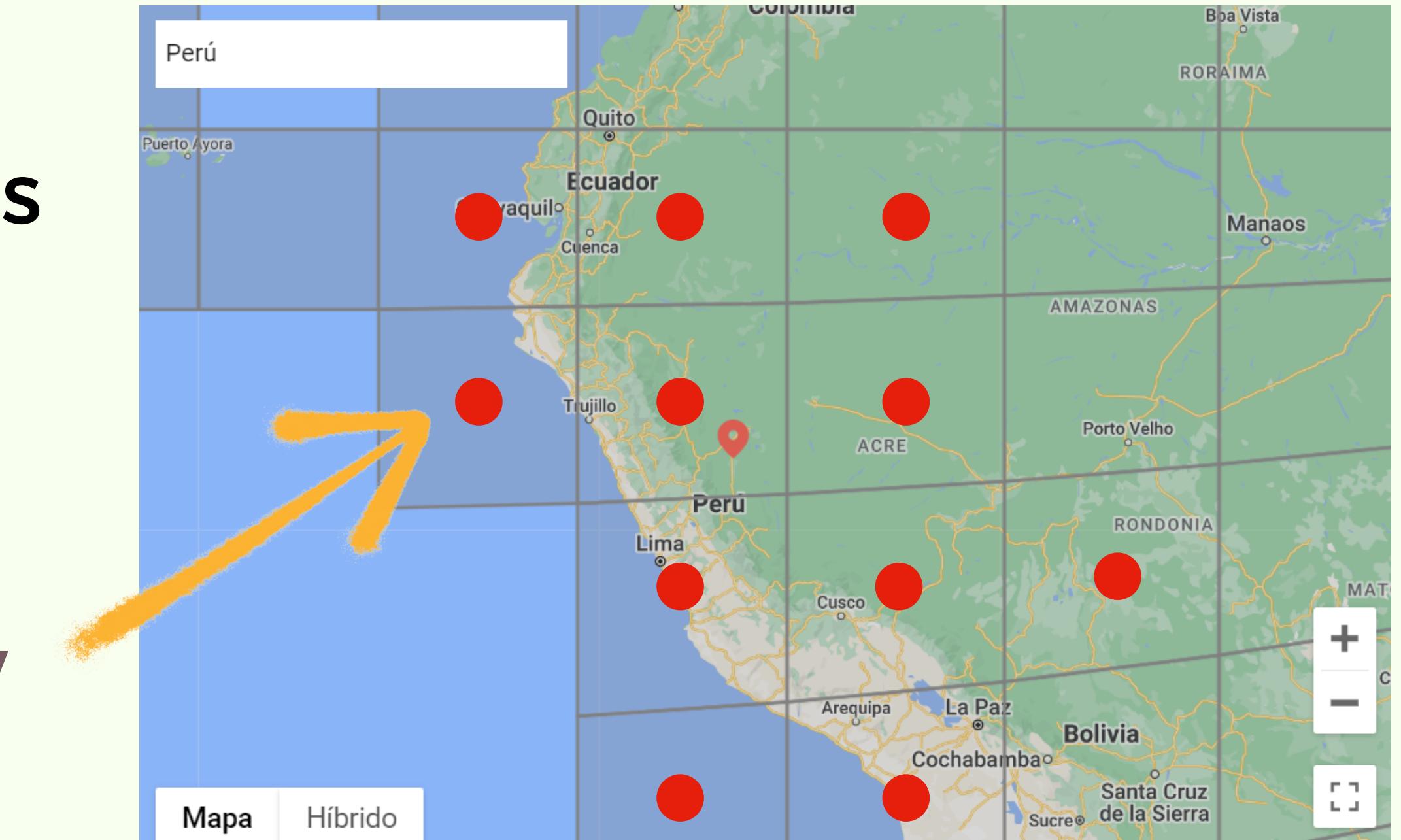
Fue generado usando un modelo de deep learning llamado U-Net.



# Open buildings

12,072,733 registros  
6 atributos

Archivo.csv



# Confidence?

---

Confianza del modelo en la habitabilidad de la edificación detectada

Toma valores en la escala de 0.5 a 1

A partir de 0.75 una edificación es considerarla habitable.

**Reflejan qué tan seguro está el modelo de que algo es un edificio.**



(a) Large buildings



(b) Complex roof structure



(c) Touching buildings



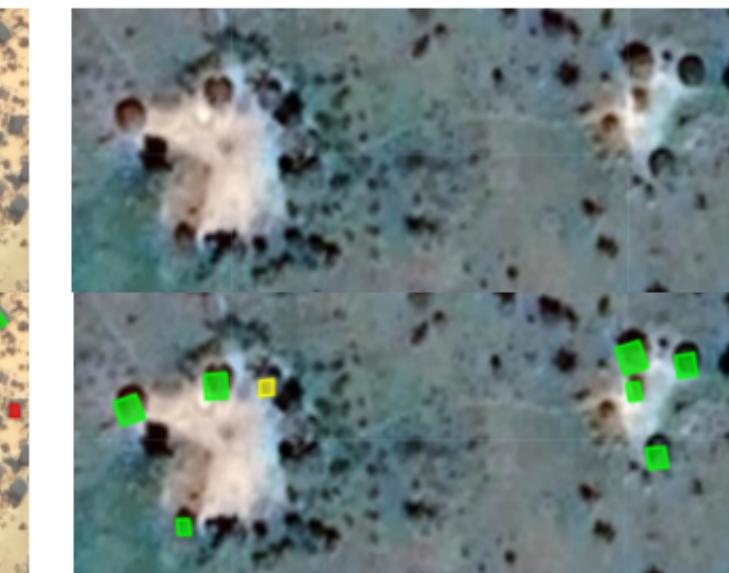
(d) Ambiguous segmentation



(e) Tree occlusion



(f) Confusing natural features



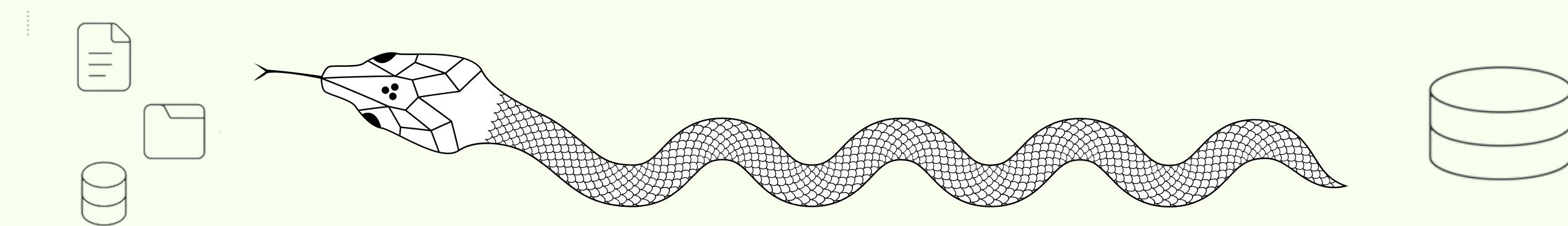
(g) Round shapes vectorized to rectangles

Potente pero no perfecto



# Preparación de los datos

- Leer y concatenar los dataframes
- Limpieza y transformación de variables cuantitativas y categóricas.
- Análisis estadístico descriptivo

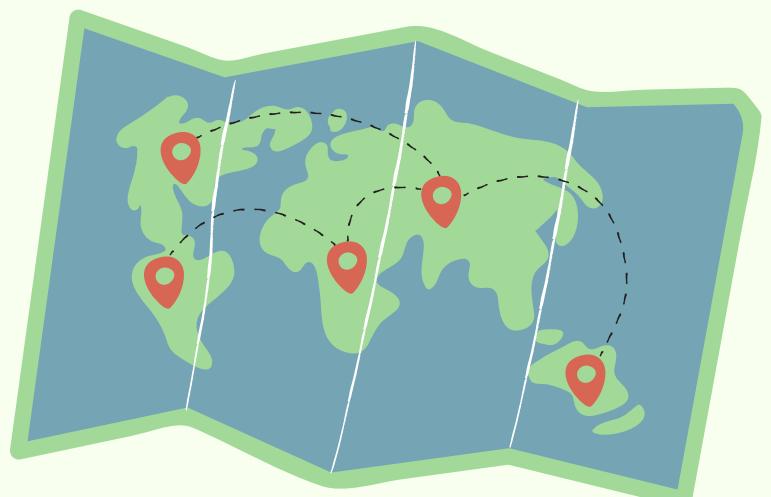
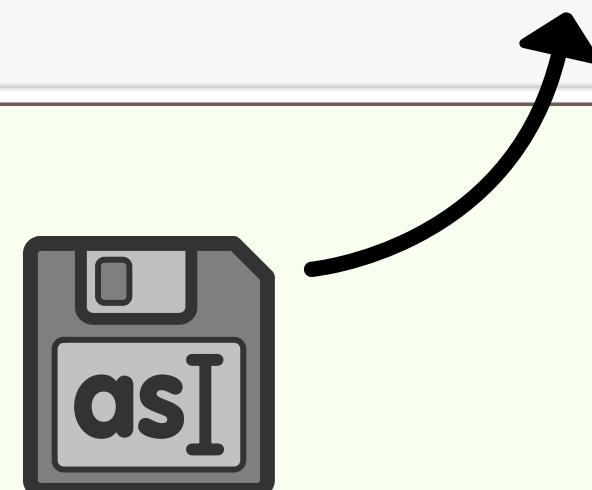


# Librerías



# Importar datos

```
df_colegios_completo = pd.read_excel('resultados_colegios.xlsx', dtype={'CODLOCAL': str})  
  
# Renombra las columnas 'NLAT_IE' y 'NLONG_IE'  
df_colegios_completo = df_colegios_completo.rename(columns={"NLAT_IE": "latitude", "NLONG_IE": "longitude"})  
df_colegios_completo.columns
```



# Datos filtrados

13

```
# Filtra y procesa los datos según tus condiciones
df_colegios_completo_listo = df_colegios_completo[
    (df_colegios_completo['D_ESTADO'] == 'Activa') &
    (df_colegios_completo['D_FORMA'] == 'Escolarizada') &
    (~df_colegios_completo['D_NIV_MOD'].str.contains('Especial|Instituto|Técnico|Escuela')) &
    (df_colegios_completo['TALUMNO'] > 0))
]
```

- Colegios **activos**
- Modalidad **escolarizada**
- No educación **especial**
- Colegios que tienen **alumnos**

## Varios Colegios en un mismo punto



1. Agrupamos colegios por latitud y longitud.
2. Unimos D\_NIV\_MOD diferentes.
3. Sumamos alumnos y docentes en cada grupo.
4. Eliminamos duplicados de latitud y longitud, conservando el primero con D\_NIV\_MOD, alumnos y docentes.
5. Eliminamos columnas de cantidad de alumnos por género, ya que no se usarán.

```
Valores de 'D_NIV_MOD' antes de la agrupación:
['Inicial - Jardín' 'Primaria' 'Secundaria' 'Inicial - Cuna-jar
dín'
 'Inicial - Cuna']

Valores de 'D_NIV_MOD' después de la agrupación:
['Inicial & Primaria' 'Primaria & Secundaria' 'Inicial' 'Primari
a'
 'Inicial & Primaria & Secundaria' 'Secundaria' 'Inicial & Secu
ndaria']
```

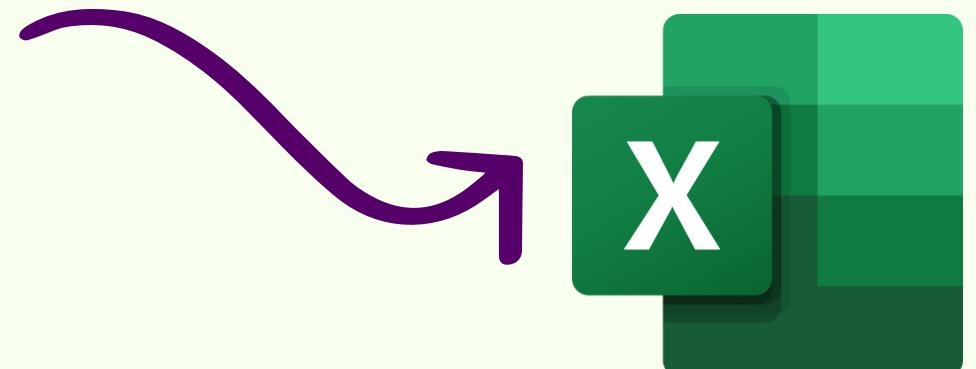
# Datos filtrados

14

Se eliminan las columnas que no usaremos

```
columnas_a_eliminar = ['Clave', 'COD_MOD', 'ANEXO', 'NIV_MOD', 'D_FORMA', 'COD_CAR',
                       'TIPSSEXO', 'GESTION', 'GES_DEP', 'DIRECTOR', 'TELEFONO', 'EMAIL',
                       'PAGWEB', 'DIR_CEN', 'REFERENCIA', 'LOCALIDAD', 'CODCP_INEI',
                       'CODCCPP', 'CEN_POB', 'AREA_CENSO', 'CODGEO', 'CODOOII',
                       'D_DREUGEL', 'TIPOPROG', 'D_TIPOPROG', 'COD_TUR', 'ESTADO',
                       'D_ESTADO', 'D_FTE_DATO', 'FECHAREG', 'FECHA_ACT']

# Eliminar las columnas del DataFrame
result_df = result_df.drop(columnas_a_eliminar, axis=1)
```



64,596 registros  
18 atributos

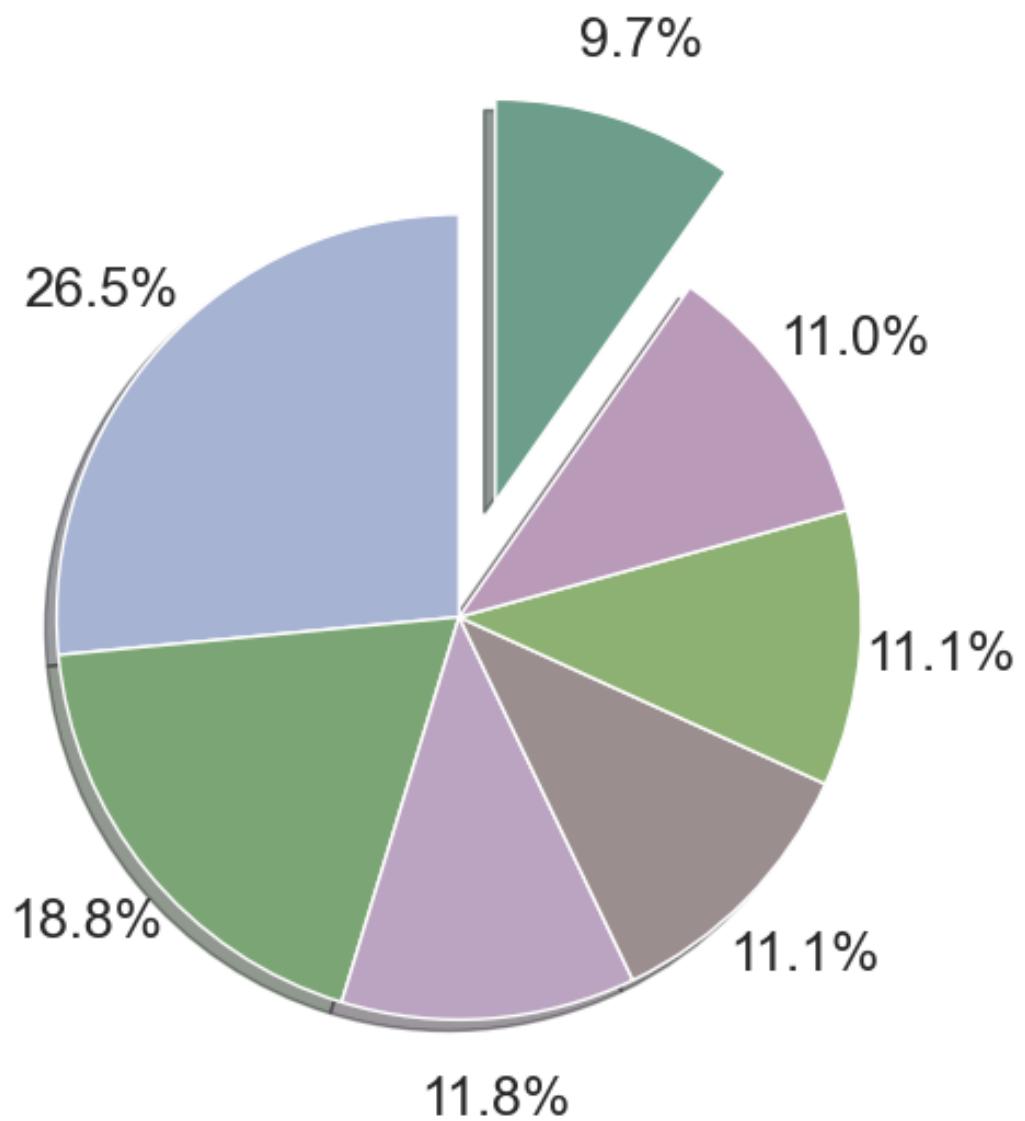
¿Qué notamos?



**DATA LISTA!**

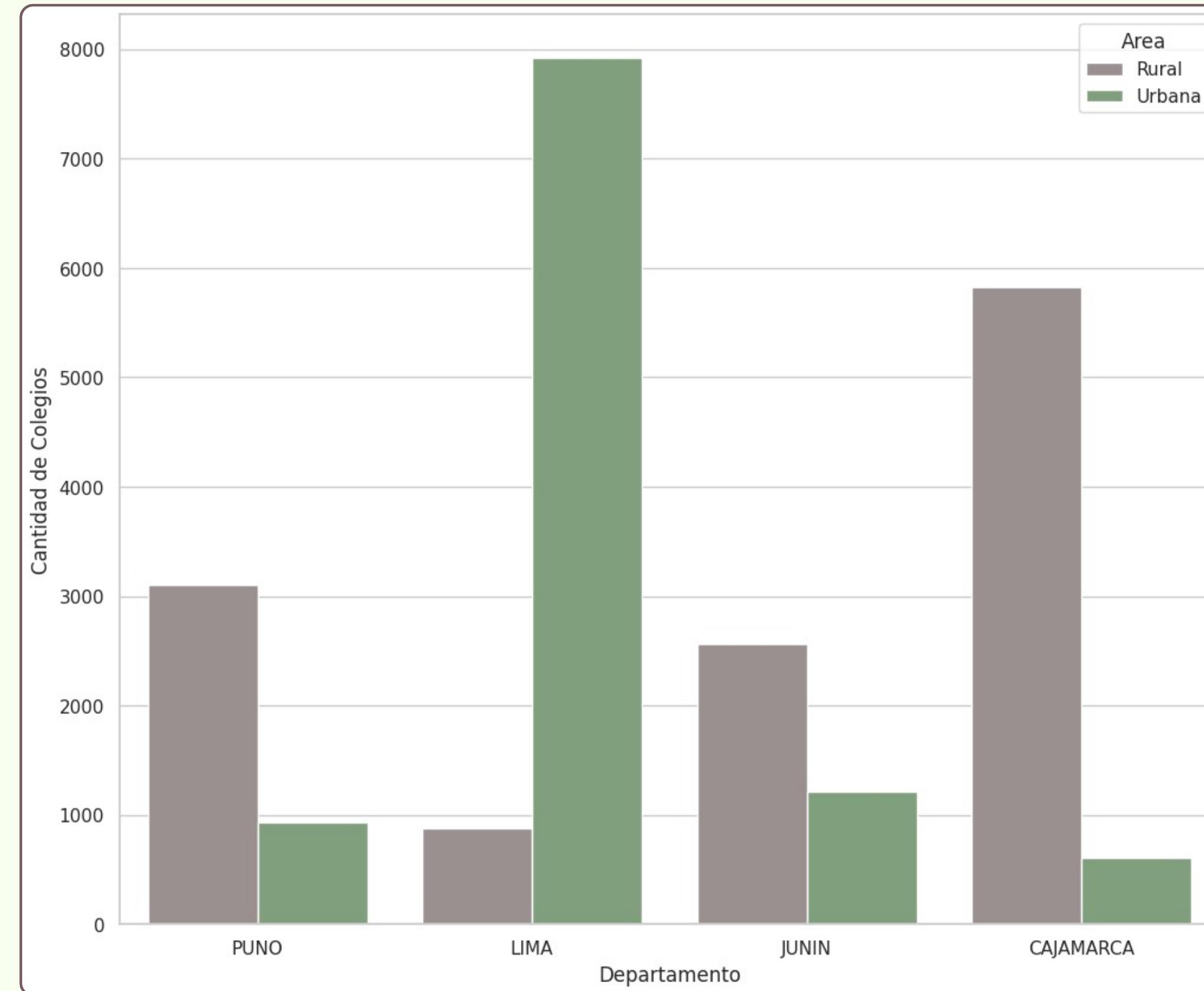
# Gráficos

Top 7 Departamentos con más colegios estatales



Se han filtrado datos que muestran los **7 departamentos** con la mayor cantidad de **colegios estatales**, y destaca **Lima** debido a su considerable población. Le sigue **Cajamarca**, que también cuenta con un número significativo de colegios. Llama la atención la presencia de **Loreto**, a pesar de tener una **población más reducida**. La lista se cierra con **Cuzco**, que representa el 9.7% de estos colegios.

# Gráficos

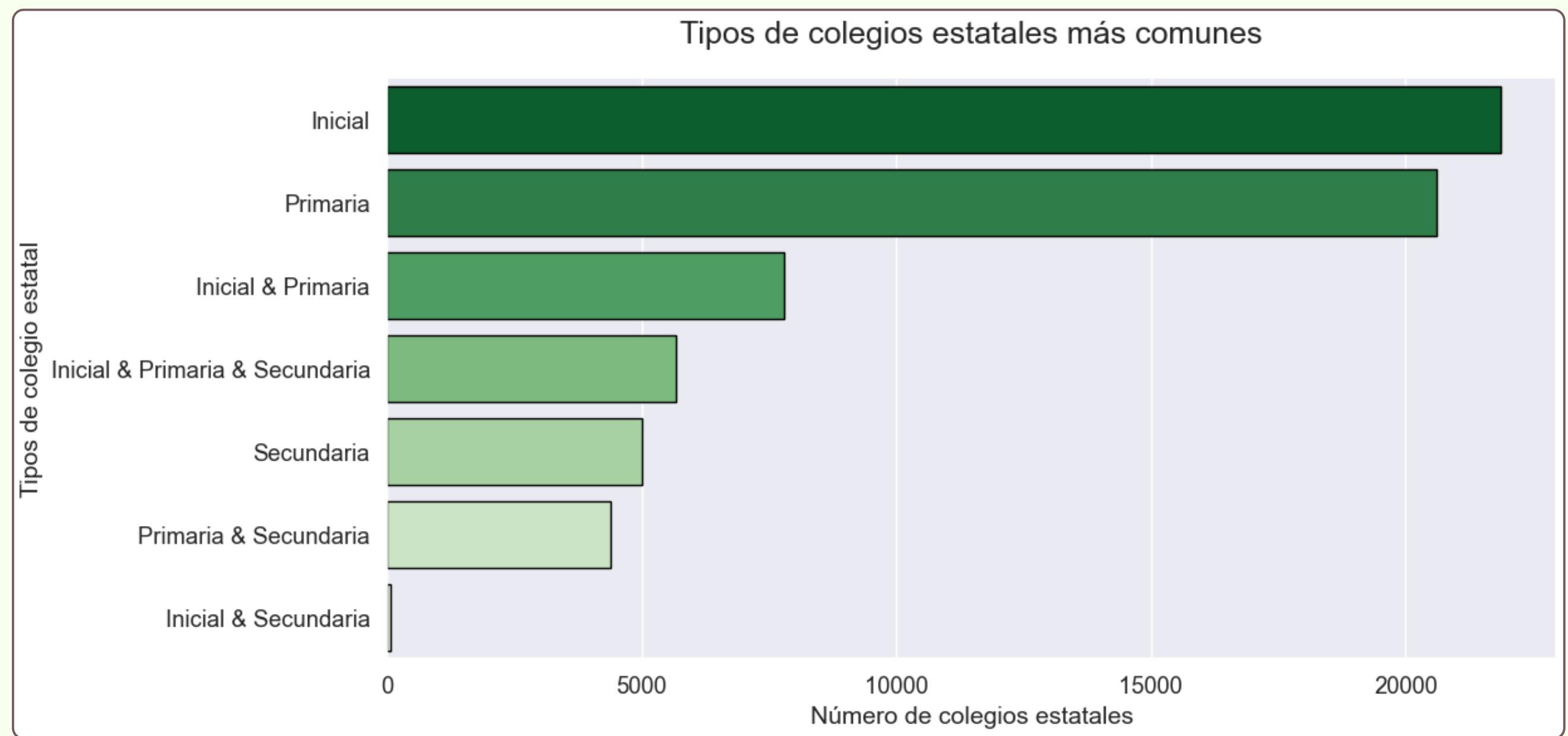


Se han filtrado datos que muestran los **4 departamentos** con la mayor cantidad de **colegios**, destaca **Lima** debido a su considerable población. Le sigue **Cajamarca**, que también cuenta con un número significativo de colegios. Algo resaltante es que todos excepto Lima cuentan con una mayoría en el **área rural**.

# Gráficos

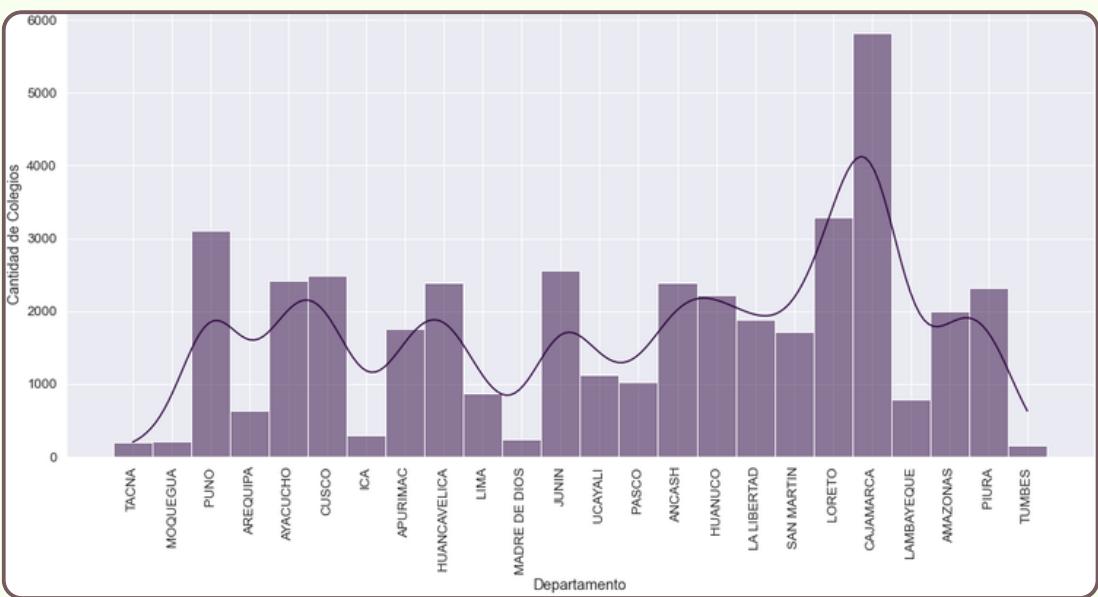
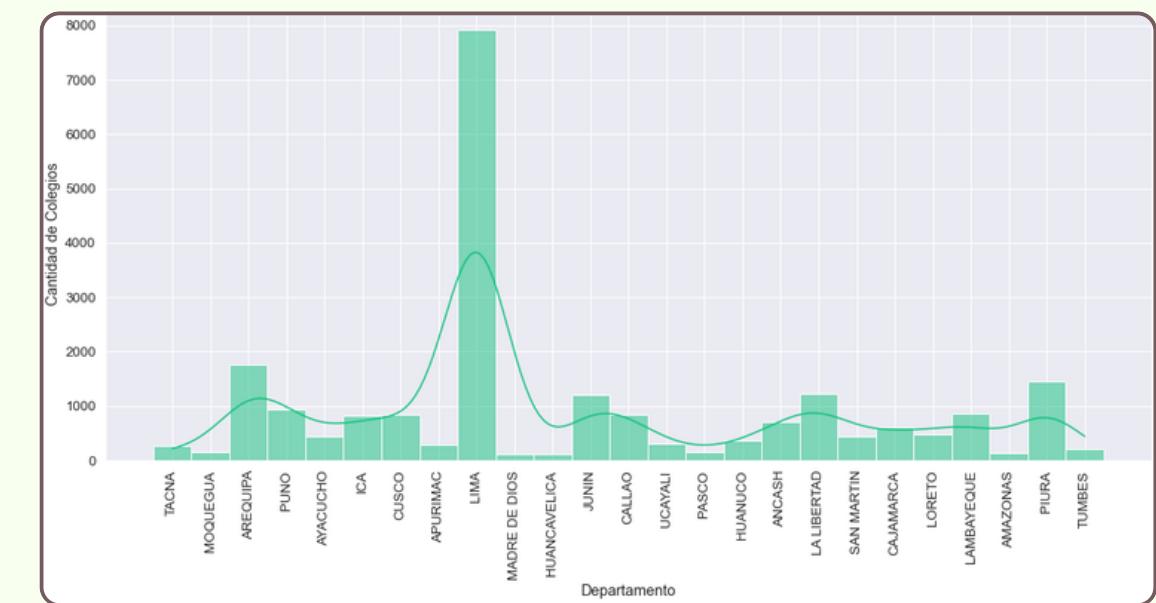
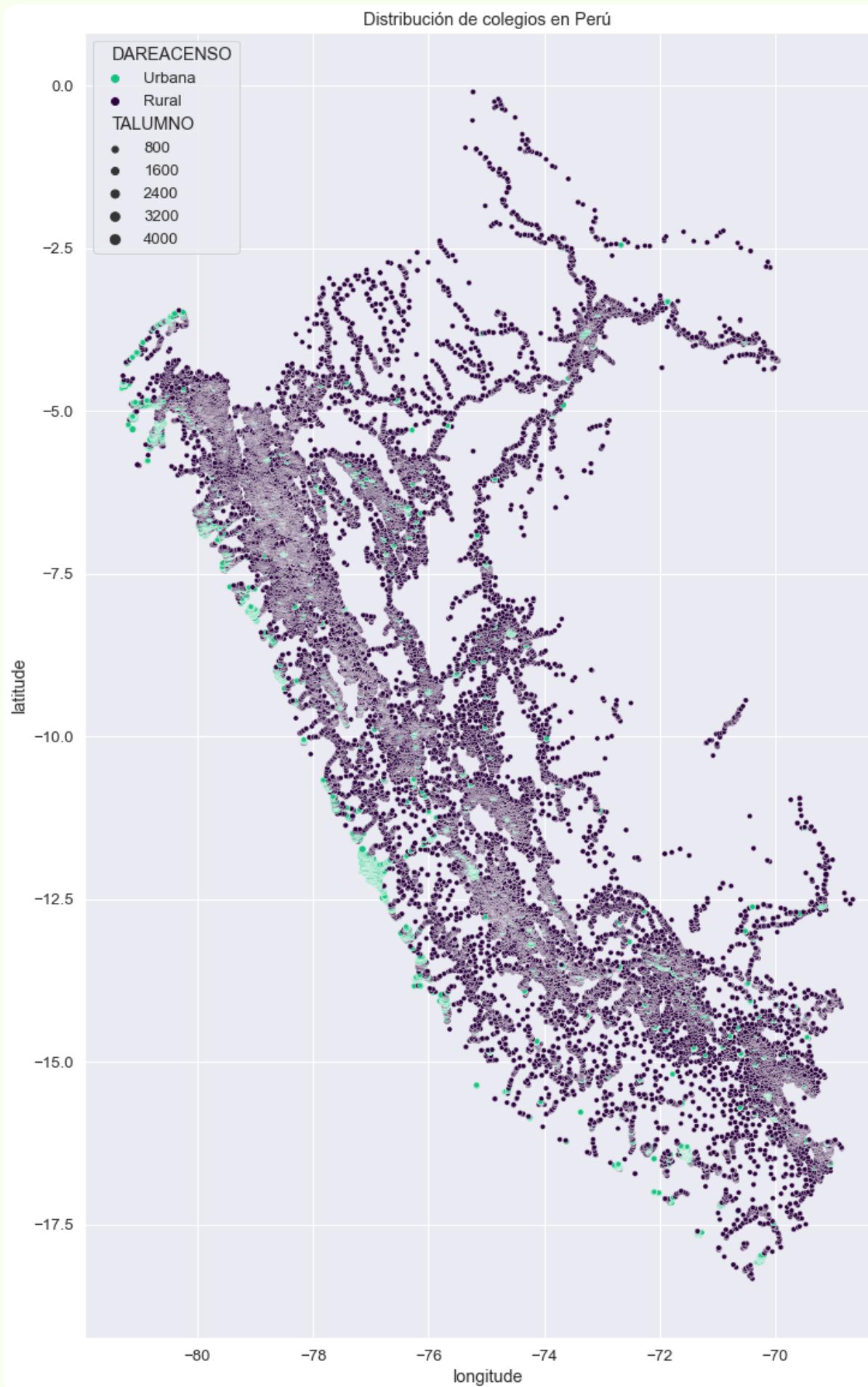
Se presenta un gráfico que exhibe los **tipos más frecuentes de colegios estatales**. La mayoría de ellos ofrecen únicamente **nivel inicial**, seguidos por aquellos que ofrecen solo educación primaria.

Aproximadamente 5000 colegios en total, alrededor del **25%** de los colegios iniciales, abarcan los **tres niveles educativos**. Los casos que ofrecen tanto inicial como secundaria son escasos, cercanos al 0%.



# Gráficos

Se presenta un mapa que representa los colegios en Perú, diferenciados entre **urbanos y rurales**, utilizando el tamaño de acuerdo con la **cantidad de alumnos** para dar una impresión de **volumen**.



Es evidente una marcada **predominancia de colegios rurales**, representados en tonos morados. También se observa una concentración significativa de **colegios urbanos a lo largo de la costa**, mientras que en otras regiones del país, estos están **dispersos y en menor cantidad**.

# Gráficos

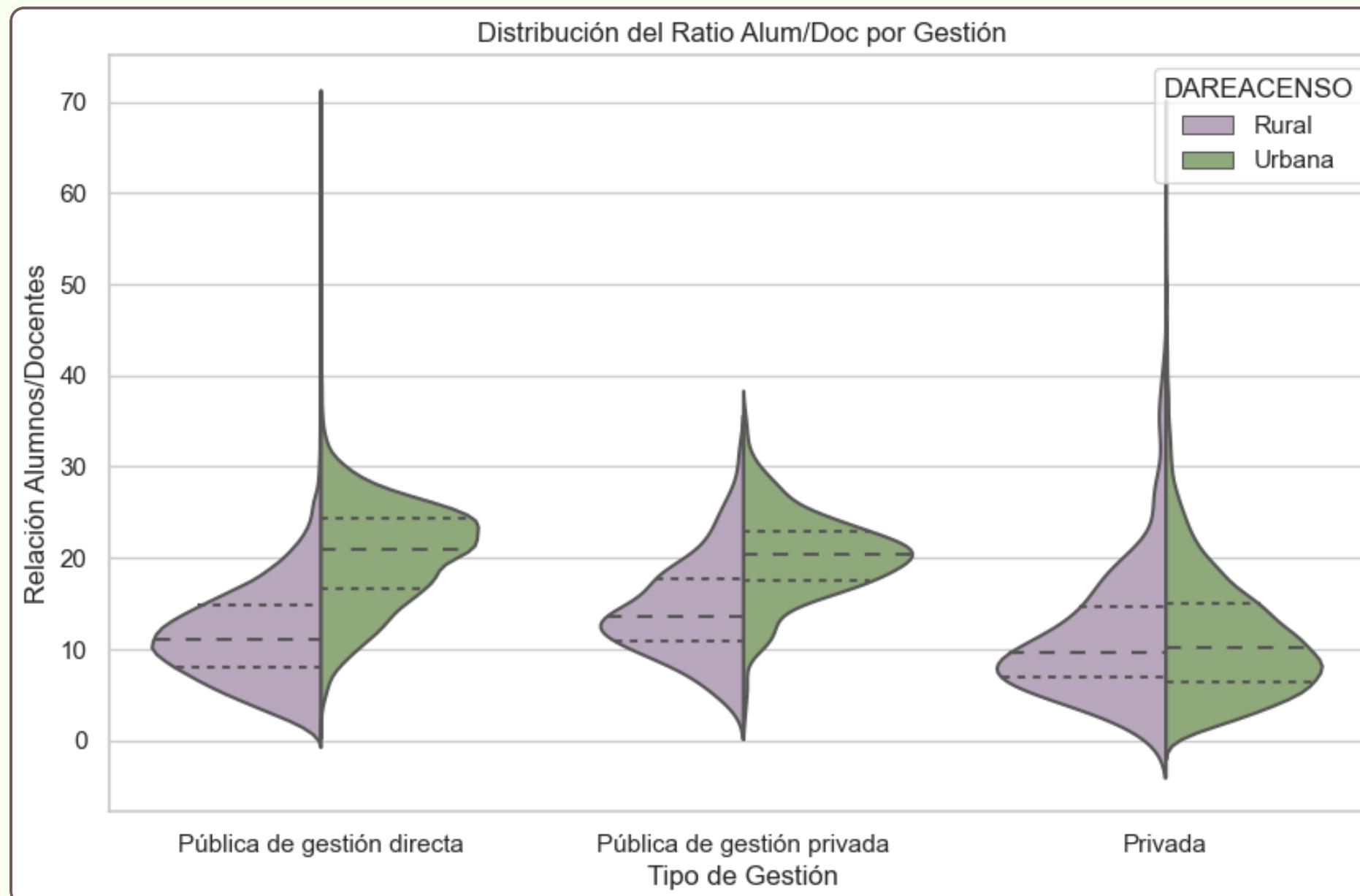
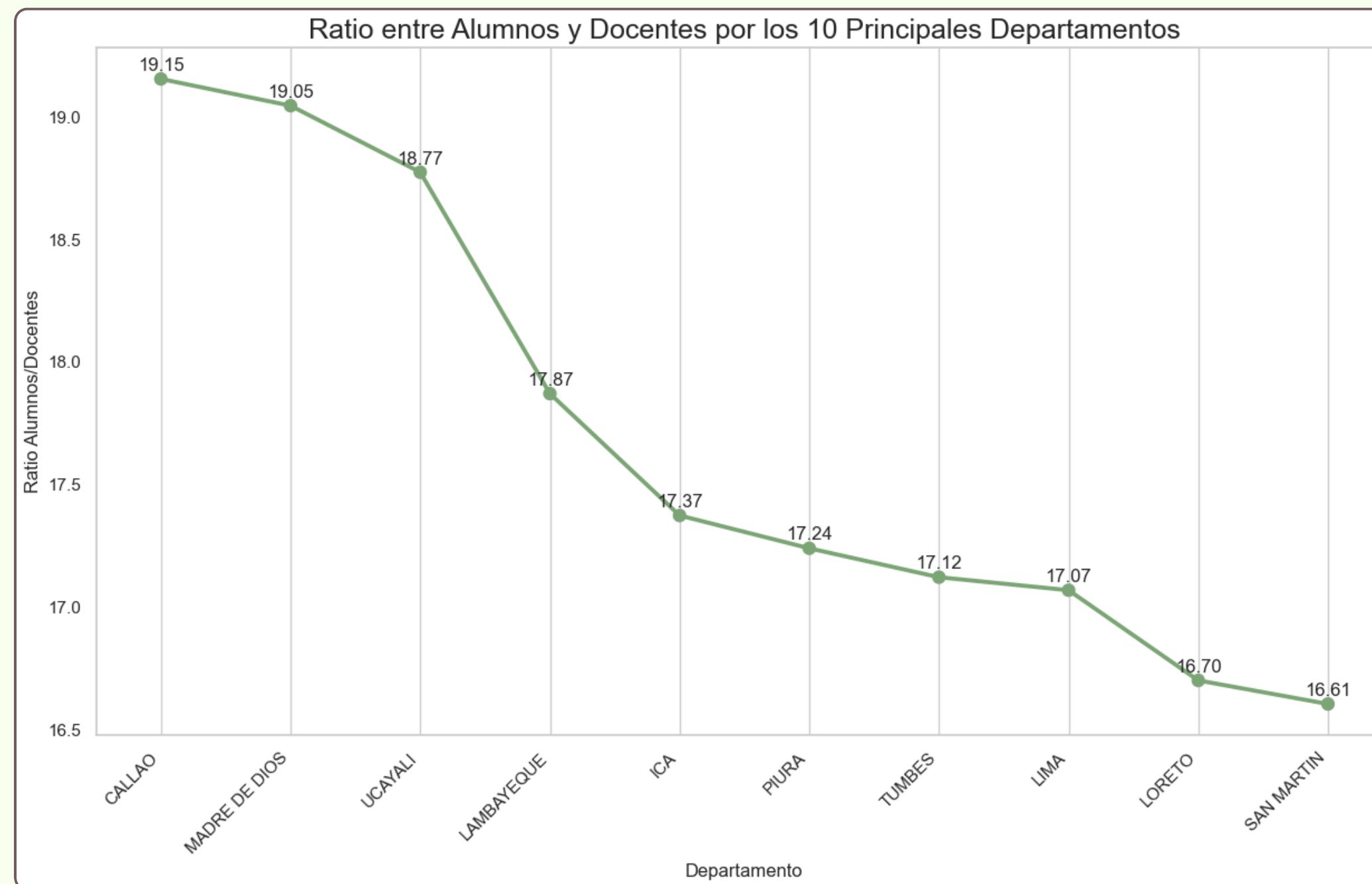


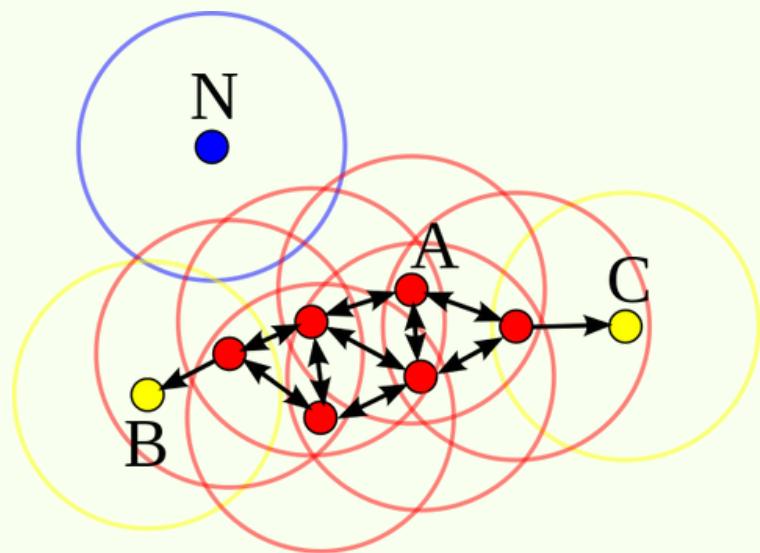
Gráfico de violín que ilustra la relación entre **alumnos y docentes** según el **tipo de gestión y el área** de censo. Observamos que los límites del ratio **son similares** en los tres grupos. Se destaca que en los colegios privados, la distribución es **casi idéntica** tanto en áreas rurales como urbanas, mientras que en los otros casos, existe una diferencia notable en el ratio en el **área urbana**.

# Gráficos



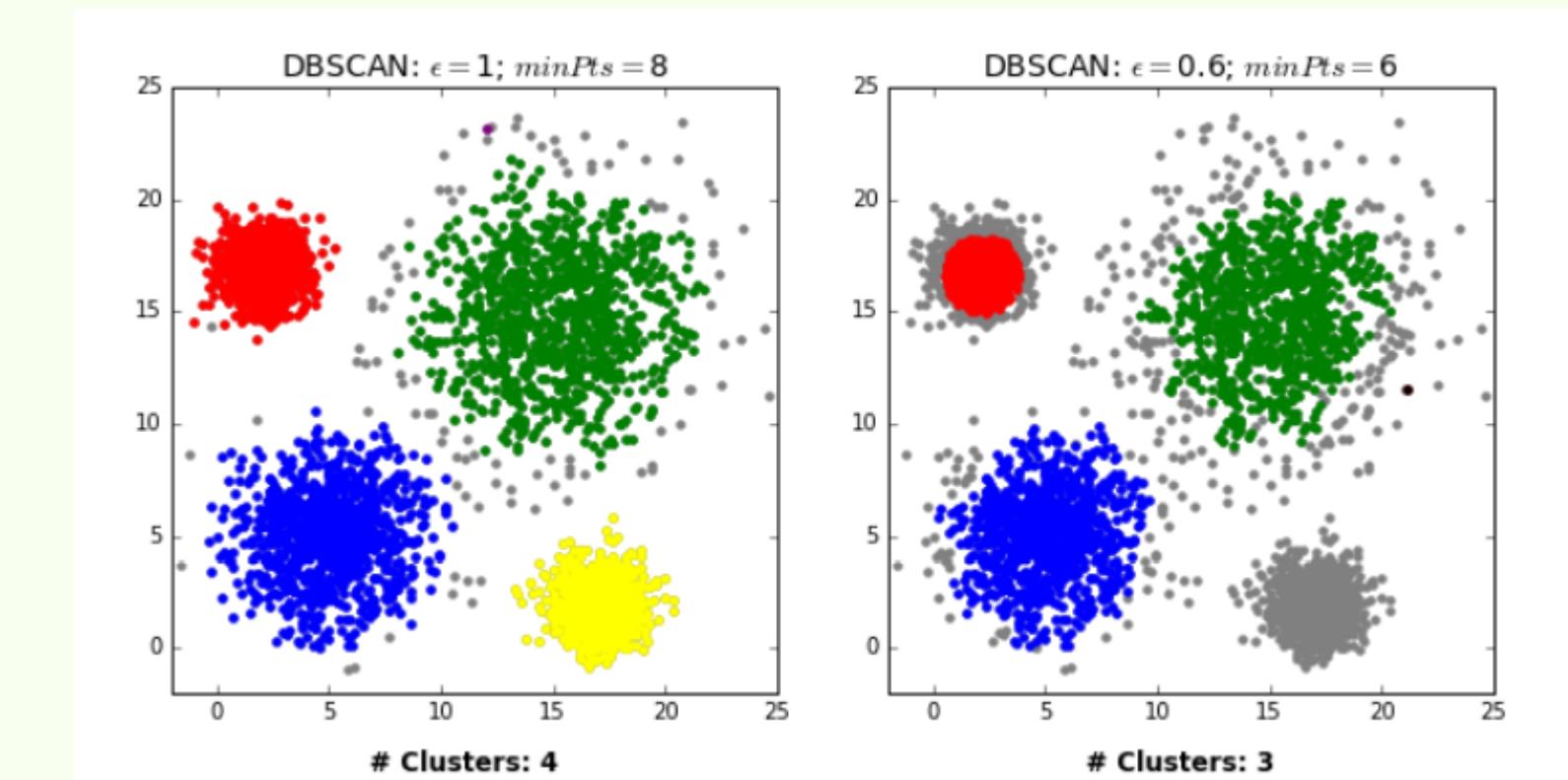
En continuación de nuestro análisis sobre el ratio entre alumnos y docentes, observamos que el departamento de **Callao** presenta el mayor ratio, con un **valor de 19.15**, una cifra notablemente elevada que podría tener un **impacto negativo** en la calidad de la educación. A Callao, le siguen departamentos de la **costa y selva**. En el top10, no se encuentra ningún departamento de la sierra.

# DBSCAN



Es un algoritmo de **clustering** que agrupa puntos de datos en función de la **densidad local** en lugar de depender de la **distancia euclíadiana** entre los puntos. Esto significa que puede encontrar clústeres de formas arbitrarias y es **resistente al ruido**.

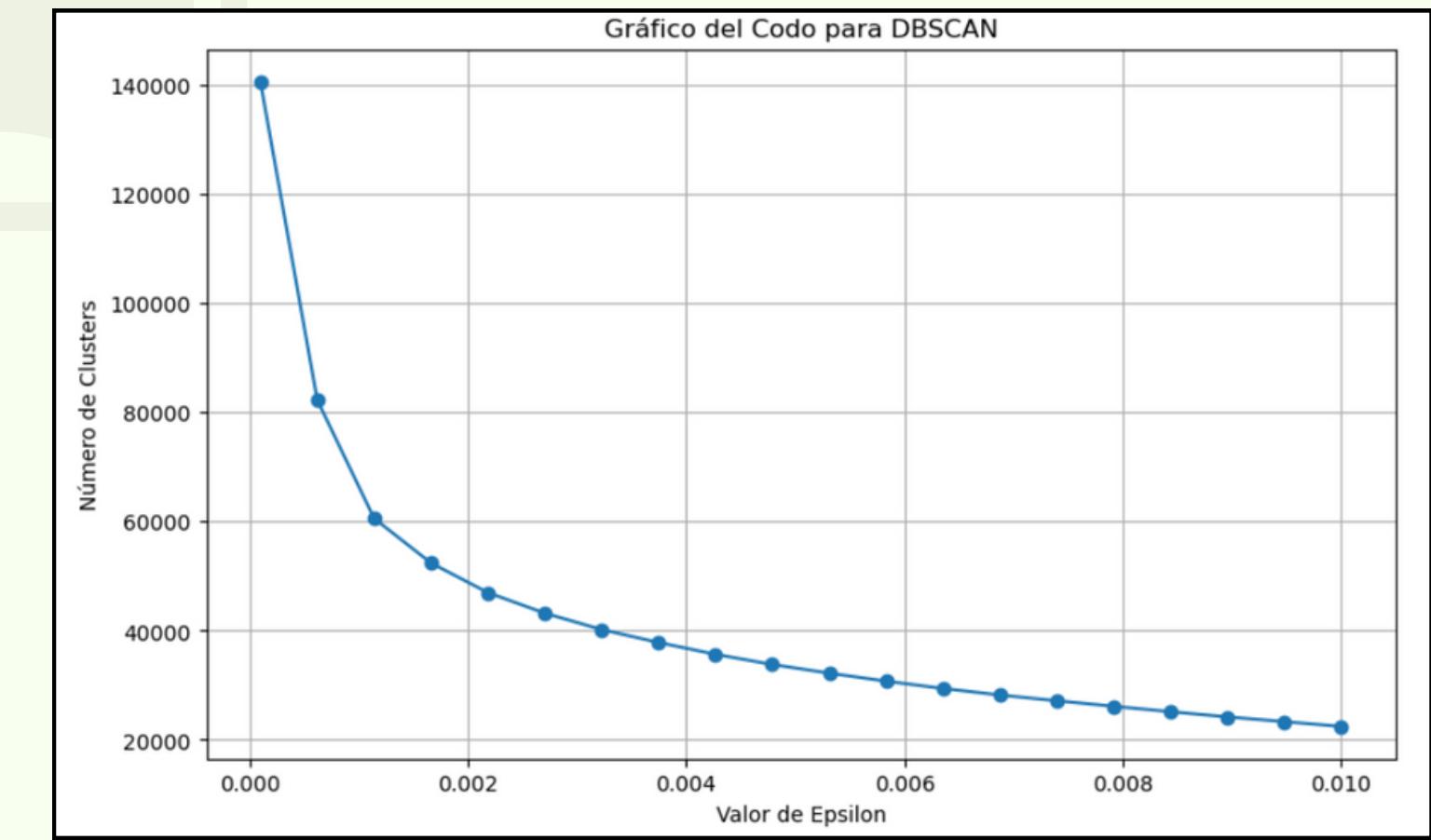
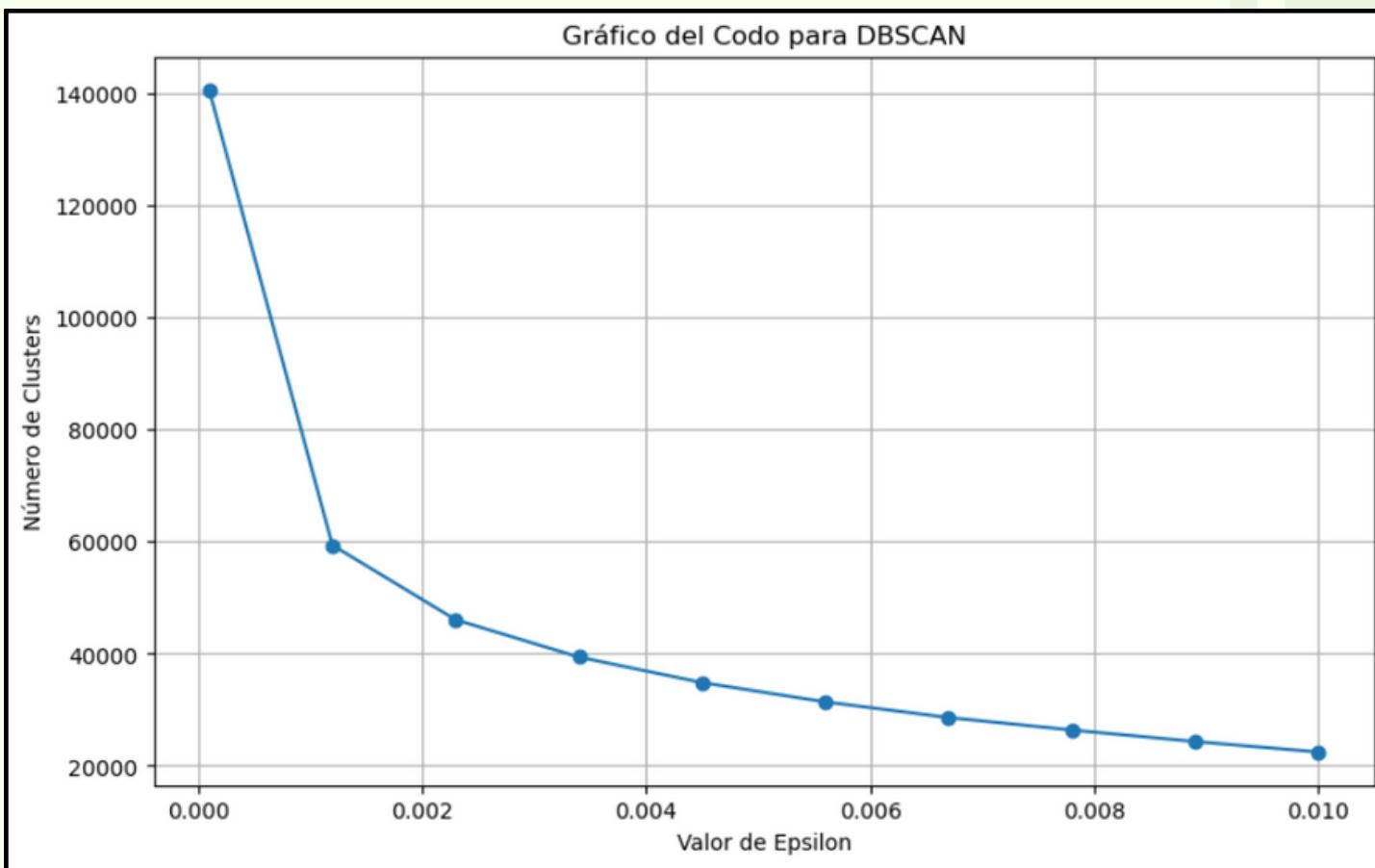
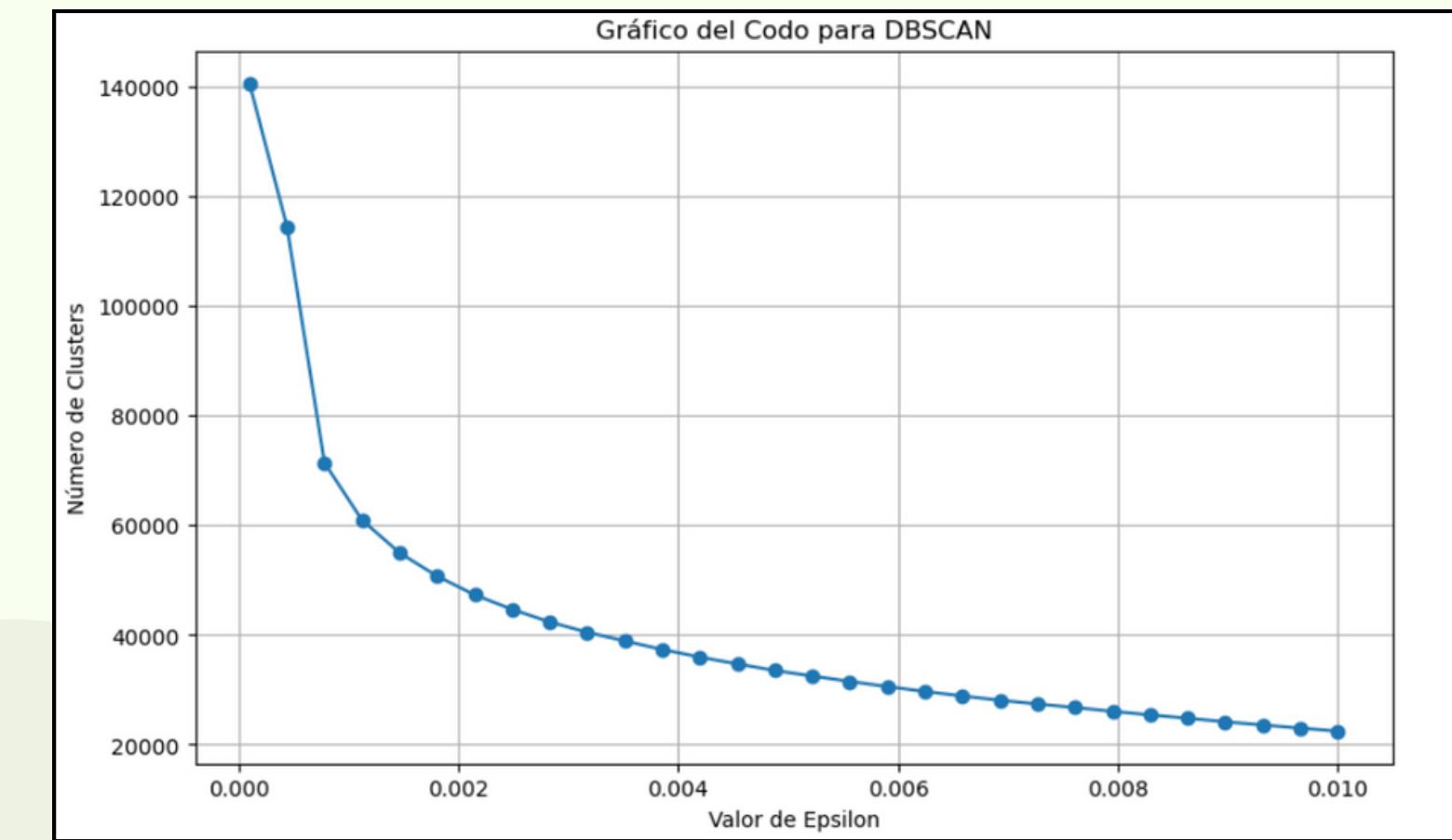
1. Selecciona un **punto de datos** no asignado.
2. Encuentra puntos cercanos dentro de un **radio  $\epsilon$**  basado en distancia.
3. Si hay al menos **MinPts puntos cercanos**, forma un clúster y etiqueta esos puntos.
4. **Repite** el proceso para expandir el clúster.
5. Continúa el proceso para puntos no asignados hasta explorar **todo el conjunto de datos**.



**Density-based spatial clustering of applications with noise**

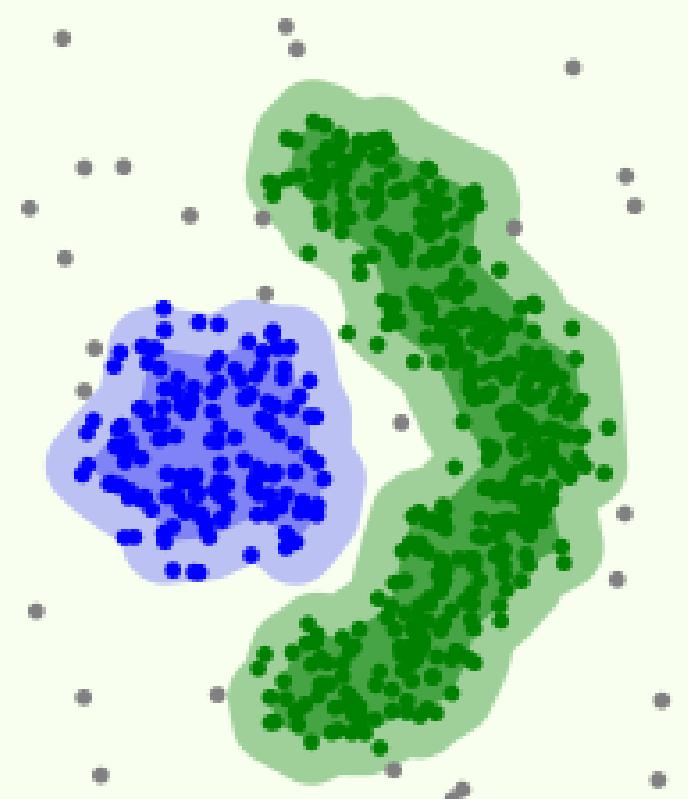
# Modelamiento

Se empleó el algoritmo DBSCAN con el propósito de encontrar el valor de epsilon ideal que permitiera maximizar la cantidad de grupos identificados. Se llevaron a cabo pruebas con diversos valores y se representó gráficamente la curva del "codo" para determinar el punto de inflexión.



# Modelamiento

Se aplicó DBSCAN con el epsilon óptimo para agrupar los puntos en clusters.



Se agregaron las etiquetas de cluster al dataframe original



**EPSILON OPTIMO  
0.00115**

	latitude	longitude	confidence	Etiqueta	cluster_label
0	-9.518850	-77.531910	NaN	1	0
1	-9.530670	-77.531960	NaN	1	1
2	-9.531100	-77.522700	NaN	1	2
3	-9.516673	-77.531481	NaN	1	0
4	-9.513940	-77.504026	NaN	1	3
...	...	...	...	...	...
1048570	-11.088639	-74.653444	0.8525	2	16231
1048571	-11.239494	-75.685245	0.7688	2	37664
1048572	-11.145996	-77.245295	0.8042	2	10360
1048573	-12.068330	-75.311771	0.7256	2	31884
1048574	-11.171449	-76.324855	0.8705	2	-1
1222282 rows × 5 columns					

# Modelamiento

Se calculó la media de "confidence" por cluster y se reemplazaron los NA por la media del cluster.

```
# Calcular la media de "confidence" por grupo
media_por_grupo = resultado.groupby('cluster_label')['confidence'].transform('mean')

# Reemplazar los valores nulos con la media por grupo
resultado['confidence'].fillna(media_por_grupo, inplace=True)

# Asegúrate de que 'confidence' sea de tipo numérico (puede haber problemas si es de tipo object)
resultado['confidence'] = resultado['confidence'].astype(float)
```



	latitude	longitude	confidence	Etiqueta	cluster_label
5	-9.535456	-77.567587	0.765876	1	-1
9	-9.514120	-77.526150	0.765876	1	-1
11	-9.516240	-77.478771	0.765876	1	-1
16	-9.502821	-77.527342	0.765876	1	-1
17	-9.599418	-77.492296	0.765876	1	-1

latitude	0
longitude	0
confidence	0
Etiqueta	0
cluster_label	0
dtype:	int64

# Modelamiento

Se obtuvieron los clusteres que tienen media de confidence menor a 0.75 (zona no habitables)

```
# Filtrar las etiquetas de clusters con una media de "confidence" menor a 0.75
cluster_means = resultado.groupby('cluster_label')['confidence'].mean()
clusters_con_media_baja = cluster_means[cluster_means < 0.75]
```

Etiquetas de clusters:  
9050

Buscamos los colegios (etiqueta=1) que se encuentren en los clusteres previamente hallados.

Índices de las filas que cumplen con ambas condiciones:

```
Int64Index([ 3803, 3828, 3844, 4720, 4721, 4723, 4727, 4730,
 4738, 4739,
 ...
 172729, 172741, 172787, 172828, 172858, 172884, 173079, 173107,
 173294, 173644],
 dtype='int64', length=1704)
```



# Modelamiento

Se extrajo información de esos colegios desde el dataframe de instituciones educativas



	Clave	COD_MOD	ANEXO	CODLOCAL	CEN_EDU	NIV_MOD	D_NIV_MOD	D_FORMA	COD_C
3803	0249607-0	249607	0	29697.0	20483	B0	Primaria	Escolarizada	
3828	0250019-0	250019	0	29942.0	20534	B0	Primaria	Escolarizada	
3844	0912485-0	912485	0	538595.0	88338	B0	Primaria	Escolarizada	
4720	0226340-0	226340	0	67528.0	SAN MARTIN DE PORRES	A2	Inicial - Jardín	Escolarizada	
4721	0226357-0	226357	0	68194.0	NIÑO JESUS DE YAUCA	A2	Inicial - Jardín	Escolarizada	

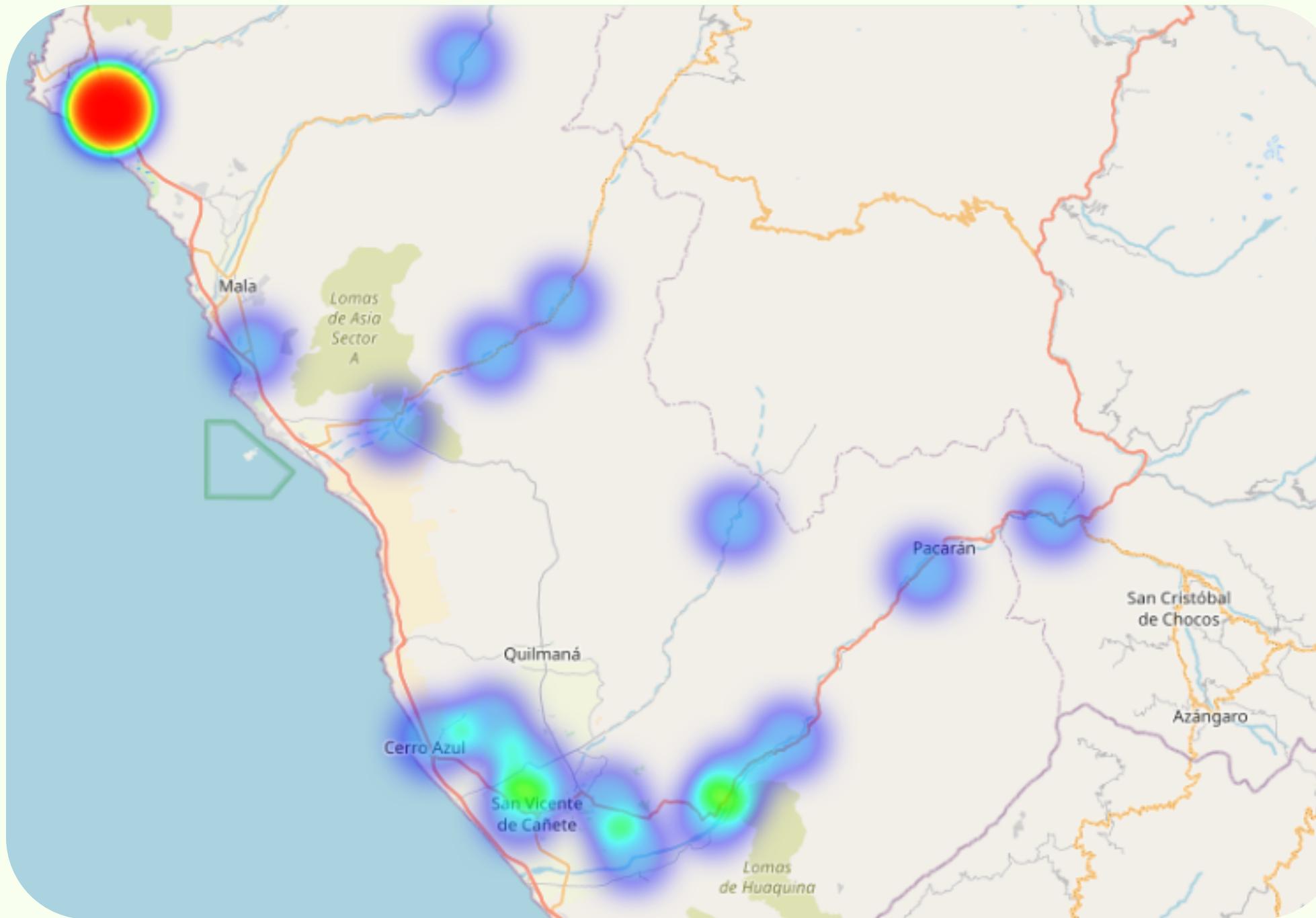
D_PROV	
CAÑETE	91
HUAROCHIRI	80
BARRANCA	39
LIMA	36
HUAURA	26
YAYOS	23
HUARAL	20
CAJATAMBO	11
OYON	10
CANTA	5

Se identificaron los departamentos y provincias con más colegios afectados.



# Modelamiento

Se graficó un mapa de calor de las coordenadas de los colegios afectados.



	latitude	longitude	confidence	Etiqueta
1	-12.257356	-76.885870	0.8953	2
4	-11.936533	-76.716992	0.7728	2
9	-12.194070	-76.986501	0.6691	2
15	-12.128998	-76.956008	0.8493	2
17	-13.000520	-76.447401	0.8348	2
...	...	...	...	...
172056	-12.628466	-76.664389	NaN	1
172087	-13.070748	-76.383584	NaN	1
172135	-12.521303	-76.737549	NaN	1
172338	-12.662970	-76.627150	NaN	1
172416	-12.949589	-76.383350	NaN	1

# Modelamiento

Se aplica DBSCAN nuevamente, esta vez solo a los datos filtrados de edificios en Cañete. Esta vez, se filtran los clusters con promedio de "confidence" mayor o igual a 0.75, para quedarse con las zonas de buena calidad.

	latitude	longitude	confidence	Etiqueta	cluster_label
27327	-13.243446	-76.265272	0.800400	2	869
27335	-12.211813	-76.837087	0.885700	2	155
27338	-12.059996	-76.921738	0.845100	2	0
27339	-12.244927	-76.857051	0.758900	2	0
27340	-12.124871	-76.998966	0.766800	2	0
...	...	...	...	...	...
170154	-12.998338	-76.449421	0.761068	1	2
171241	-13.074650	-76.323854	0.765964	1	8
171242	-12.657762	-76.632363	0.760127	1	41
172055	-12.542154	-76.721523	0.763744	1	90
172416	-12.949589	-76.383350	0.767648	1	20

[554 rows x 5 columns]



# Modelamiento

Se eliminan los clusters que contengan colegios (etiqueta 1), para quedarse solo con clusters de edificios, donde se podría reubicar un colegio.

Se calculan los centroides de cada cluster de edificios.

Para cada colegio previamente identificado como mal ubicado, se calculan las distancias a los centroides de los clusters y se muestran los 2 más cercanos, como posibles nuevas ubicaciones.

Escuela en Cañete - Coordenadas: (-12.5198, -76.7377)

Centroide 103 - Coordenadas: (-12.52787038, -76.74191334) - Distancia: 0.00910402478577083

Centroide 28 - Coordenadas: (-12.542064666666667, -76.72149433333334) - Distancia: 0.027537955877094625

Escuela en Cañete - Coordenadas: (-13.078289, -76.244286)

Centroide 102 - Coordenadas: (-13.11289, -76.27011) - Distancia: 0.043175319072359226

Centroide 8 - Coordenadas: (-13.07854775, -76.28841800000001) - Distancia: 0.04413275853108343

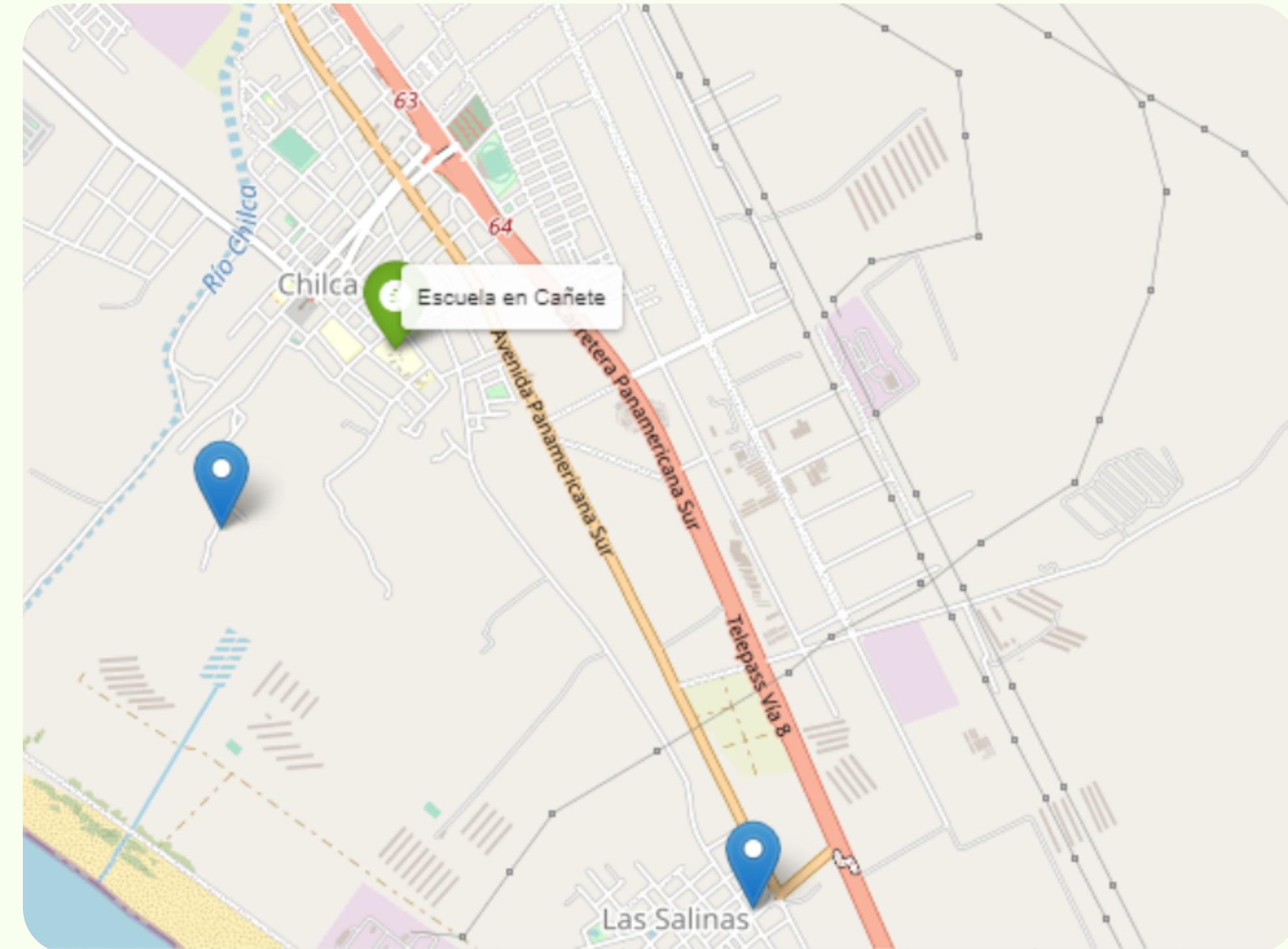
Escuela en Cañete - Coordenadas: (-12.76893, -76.503857)

Centroide 63 - Coordenadas: (-12.747542, -76.493563) - Distancia: 0.023736321956024178

Centroide 46 - Coordenadas: (-12.779398666666665, -76.555812) - Distancia: 0.0529991981710899

# Modelamiento

Se grafica un mapa interactivo con un marcador para la escuela y marcadores para los 2 centroides más cercanos, como ejemplo visual.



# PREGUNTAS

