Análisis de la distribución de centros educativos en el Perú: un enfoque de recomendación para la formulación de políticas educativas

Juan Carlos Lindo Mercado
Ingeniería de la Información
jc.lindom@alum.up.edu.pe

Sergio Andrés Martinez Vasquez
Ingeniería de la Información
sa.martinezva@alum.up.edu.pe

Franz Figueroa Guaylupo
Ingeniería de la Información
f.figueroag@alum.up.edu.pe

Jose Carlos Salinas Málaga
Ingeniería Empresarial
jc.salinasma@alum.up.edu.pe

***Abstract*—The concerning prevalence of inadequate school infrastructure and inequitable access to quality education in Peru prompted this project to analyze the location and distribution of schools using data mining algorithms. By leveraging a comprehensive dataset from the Ministry of Education about educational institutions supplemented with satellite imagery of buildings from Open Buildings, the project aims to develop an algorithm to identify schools with poor surrounding infrastructure. The overarching goal is to provide data-driven insights and generate an actionable report for the Ministry of Education, highlighting legislative proposals based on the analysis to potentially improve educational access, transportation options, infrastructure safety, and learning environments. Rather than prescribing specific policies, this project seeks to supply key analytics to inform the government's decision-making process regarding investments and interventions needed to promote equitable access to quality education across Peru.**

***Index Terms*—education, recommendation model, relocation, DBSCAN**

I. Introducción

La educación, tal como Hanushek y Woeman [1] han brillantemente planteado, ejerce un rol determinante en el crecimiento académico, pues el aprendizaje ostenta un profundo impacto en la construcción del capital humano. Glewwe y Kremer [2], por su parte, exploran la relevancia de la educación desde la perspectiva del bienestar social. En aras de sustentar sus investigaciones, postulan una función de aprendizaje que gravita en torno a múltiples factores intrínsecos al estudiante, al docente y al entorno escolar donde se desenvuelve este delicado proceso.

La búsqueda constante de la mejora en la calidad de las instituciones educativas no es un fenómeno exclusivo de Perú, sino una corriente global, como lo atestigua Coleman et al. [3]. En dicho estudio, se empeñan en predecir el éxito académico de los estudiantes estadounidenses, basándose en variables socioeconómicas y escolares. Curiosamente, los hallazgos arrojaron que las variables socioeconómicas ostentaban la mayor relevancia. Sin embargo, en un giro de los acontecimientos, el Banco Mundial [4], en 1990, presentó evidencia que contradice la tesis planteada por Coleman en 1966, generando un intrigante debate.

Siguiendo el hilo argumental propuesto por el Banco Mundial [4], Fertig y Schmidt [5] y Rothstein [6] sugieren que los determinantes que inciden en la calidad del servicio educativo se pueden dividir en dos amplios grupos: los factores relacionados con la oferta y los factores vinculados a la demanda. En esta ocasión, nos centraremos en los determinantes de la demanda. Según estos autores, dichos determinantes abordan la cobertura del servicio y el ambiente donde se imparte la educación. Trasladando esta noción a la realidad peruana, según un informe de la UNESCO de 2019, las instituciones educativas que carecen de servicios esenciales, como instalaciones sanitarias adecuadas, pizarras en condiciones óptimas y mobiliario adecuado, se asocian inequívocamente con una educación de calidad inferior.

Los estudios previos subrayan la importancia del entorno donde se presta el servicio educativo, especialmente en términos de infraestructura. No obstante, la infraestructura de las escuelas no es el único factor a considerar, ya que los niños y, sobretodo, los adolescentes, son altamente susceptibles a diversos estímulos visuales que pueden incitar comportamientos inadecuados a largo plazo. Todo esto halla apoyo en la sólida base de la teoría psicológica del conductismo.

El propósito fundamental de esta investigación reside en la identificación de instituciones educativas que operan en contextos circundados por infraestructuras deficientes. Para alcanzar tal objetivo, emplearemos un enfoque respaldado por dos fuentes de información clave: el conjunto de datos proporcionado por el Ministerio de Educación y un conjunto complementario de datos geoespaciales de Open Buildings, que se compone de imágenes satelitales. A través de este análisis, no solo evaluaremos la calidad de la infraestructura escolar, sino también exploraremos la interacción potencial

entre este entorno físico y el comportamiento a largo plazo de los estudiantes.

La siguiente fase de nuestra investigación se centrará en la evaluación de la viabilidad de trasladar aquellas instituciones educativas ubicadas en áreas con infraestructura deficiente hacia entornos con infraestructura más adecuada. Este proceso no solo considerará la calidad de la infraestructura, sino también la optimización de la ubicación de estos nuevos colegios, de manera que se facilite el acceso de todos los niños a las escuelas, reduciendo el tiempo de viaje que actualmente implica hasta dos horas. Los resultados de esta investigación no solo proporcionarán una valiosa fuente de información, sino que también servirán como una hoja de ruta precisa y efectiva para el Ministerio de Educación. A partir de estos hallazgos, se podrán diseñar e implementar medidas concretas destinadas a la mejora y optimización integral del sistema educativo en nuestro país.

II. Motivación

Nuestra motivación principal es identificar colegios que estén ubicados en zonas inapropiadas para el desarrollo de los alumnos, y luego reubicarlos en áreas más adecuadas. Esto se hace con el fin de garantizar que la calidad de la educación no se vea perjudicada por problemas de infraestructura. Nuestro objetivo es mejorar integralmente los centros educativos y ofrecer un servicio educativo de mayor calidad respaldado por un entorno de estudio propicio y tranquilo.

Es preocupante observar que solo el 30% de la infraestructura en áreas rurales está en condiciones óptimas, segun el Ministerio de Educación (MINEDU), lo que representa un problema significativo en la calidad de las instalaciones escolares. Además, algunos estudiantes deben caminar durante una o incluso tres horas para llegar a escuelas pequeñas, lo que disuade a muchos de asistir regularmente a clases debido a los riesgos asociados con este trayecto. En este contexto, es evidente la necesidad de abordar estas cuestiones y proponer soluciones a través de un ensayo académico. Esto es fundamental para mejorar el sistema educativo peruano y asegurar un acceso equitativo a la educación para todos los estudiantes. Nuestro proyecto busca generar informes para el Ministerio de Educación con propuestas de ley para mejorar las condiciones educativas que enfrentan muchos estudiantes.

III. Preguntas de investigación y objetivos

A. Objetivos generales

1) Proveer análisis de datos y recomendaciones al Ministerio de Educación para mejorar el acceso, infraestructura y calidad de la educación en Perú.
2) Informar el proceso de toma de decisiones del gobierno sobre inversiones necesarias en el sistema educativo peruano.

B. Objetivos específicos

1) Identificar escuelas en zonas con infraestructura deficiente mediante análisis de datos.
2) Generar visualizaciones y mapas que resalten disparidades geográficas en el acceso a educación de calidad.
3) Modelar escenarios de reubicación considerando cercanía a estudiantes y mejora en condiciones de infraestructura.
4) Utilizar técnicas de minería de datos como DBSCAN para agrupar y analizar los datos.

C. Preguntas de investigación

1) ¿Es posible identificar ubicaciones óptimas para los centros educativos en Lima aplicando algoritmos de geolocalización y sistemas de información geográfica, de modo que se mejore la distribución de la oferta educativa?
2) ¿Es posible prever los principales retos logísticos, sociales y culturales en la implementación de la reubicación de centros educativos en Lima, mediante un análisis integral de factibilidad del proyecto?

IV. Recopilación y Preparación de datos

Para el desarrollo de este proyecto se utilizaron dos fuentes de datos. La primera es un dataset con la relación de las instituciones y programas educativos a nivel nacional [7], recuperado del Repositorio de Datos Abiertos del Ministerio de Educación. Y la segunda es un dataset de edificaciones de Africa, Asia, América Latina y el Caribe, resultado del proyecto Open Buildings de Google Research [8]. Para la creación de este dataset, se implementó un modelo de deep learning llamado U-Net que permitió identificar y catalogar las huellas de los edificios a partir de imágenes de satélite de alta resolución.

El dataset que comprende las instituciones educativas se compone de un total de 173,708 registros, abarcando 51 atributos. No obstante, en aras de cumplir con los objetivos de este proyecto, se llevó a cabo un proceso de filtrado que resultó en la selección de un subconjunto más enfocado. A continuación, se describen los criterios aplicados en dicho proceso:

1) **Filtrado de Instituciones Activas:** En primer lugar, se procedió a descartar aquellas instituciones educativas que no se encuentran en estado activo, lo cual asegura que los datos seleccionados estén relacionados con instituciones que están actualmente operativas.
2) **Selección de Instituciones Escolarizadas:** Se efectuó un filtro adicional para incluir únicamente aquellas instituciones que ofrecen educación escolarizada, excluyendo otras modalidades educativas.
3) **Filtrado por Niveles Educativos:** Se optó por mantener en el conjunto de datos solo las instituciones

pertenecientes a los niveles educativos de Inicial, Primaria y Secundaria, eliminando las demás categorías educativas.

4) **Eliminación de Instituciones con 0 Alumnos:** Se excluyeron las instituciones que presentaban un número total de alumnos igual a cero, garantizando que los datos recopilados reflejen contextos educativos con población estudiantil.

Además, durante el análisis de los datos, se observó que algunos registros correspondían a diferentes niveles de enseñanza en una misma institución educativa. Por consiguiente, se llevó a cabo un proceso de agrupación de estas instituciones, considerando las siguientes combinaciones: Inicial y Primaria, Primaria y Secundaria, Inicial y Secundaria, así como Inicial, Primaria y Secundaria.

Por último, se seleccionaron 18 de los 51 atributos pues son los mas relevantes para el analisis y modelamiento posterior. Los atributos del dataset final son:

- **CODLOCAL:** Código de identificación del local educativo.
- **CEN_EDU:** Número y/o nombre del servicio educativo.
- **D_NIV_MOD:** Nivel o modalidad educativa (Inicial, Primaria, Secundaria, etc.).
- **D_COD_CAR:** (Este atributo se menciona pero no se proporciona una descripción. Puede ser necesario aclarar su significado).
- **D_TIPSSEXO:** Género de los alumnos (por ejemplo, Masculino, Femenino, Mixto).
- **D_GESTION:** Tipo de gestión del servicio educativo (pública, privada, etc.).
- **D_GES_DEP:** Dependencia de la gestión del servicio educativo (puede indicar si es del Gobierno Central, Gobierno Regional, Local, etc.).
- **DAREACENSO:** Detalle del área geográfica calculada para el Censo Educativo (Censo educativo 2017).
- **D_DPTO:** Nombre del departamento donde está ubicado el local educativo.
- **D_PROV:** Nombre de la provincia donde está ubicado el local educativo.
- **D_DIST:** Nombre del distrito donde está ubicado el local educativo.
- **D_REGION:** Dirección o Gerencia regional de educación que administra la institución educativa.
- **LATITUDE:** Coordenada geográfica de latitud donde está ubicado el local educativo.
- **LONGITUDE:** Coordenada geográfica de longitud donde está ubicado el local educativo.
- **D_COD_TUR:** Detalle del turno de atención (por ejemplo, Mañana, Tarde, Noche, etc.).
- **TALUMNO:** Total de alumnos que estudian en la institución educativa.
- **TDOCENTE:** Total de docentes que enseñan en la institución educativa.
- **TSECCION:** Total de secciones existentes en la institución educativa.

Finalmente, el dataset resultante consta de 64,596 registros y está compuesto por 18 atributos.

En cuanto al dataset proporcionado por Open Buildings, se trató de una fuente de datos global que requirió ser adaptada a las necesidades del proyecto específico centrado en el contexto peruano. Inicialmente, se obtuvo un extenso conjunto de datos que constaba de 12,072,733 registros, y se incluían seis atributos.

Después de un proceso de selección y limitación, se decidió focalizar la atención en tres atributos clave que se consideraron esenciales para los objetivos de análisis:

- **LATITUDE:** Latitud del centroide del edificio.
- **LONGITUDE:** Longitud del centroide del edificio.
- **CONFIDENCE:** Confianza del modelo en la habitabilidad de la edificación detectada, valorándola en una escala de 0.5 a 1, con un umbral de 0.75 o superior para considerarla habitable.

Estos procedimientos garantizan que los datos resultantes sean relevantes y apropiados para el análisis y la extracción de conocimiento. Como resultado, se dispone de un conjunto de datos más preciso y útil para abordar con eficacia los objetivos del proyecto.

V. Exploración de Datos

En el contexto de la exploración de datos, esta sección representa una etapa inicial y crucial en nuestra investigación. Aquí, desplegamos gráficos que revelan información de sustancial relevancia, cuya comprensión es fundamental para los subsiguientes análisis y la construcción de modelos de datos.

El propósito fundamental de esta fase radica en el entendimiento de la estructura de nuestros datos, así como en la identificación de las variables que encierran un valor distintivo y que, por consiguiente, se convierten en las piedras angulares de nuestro enfoque analítico.

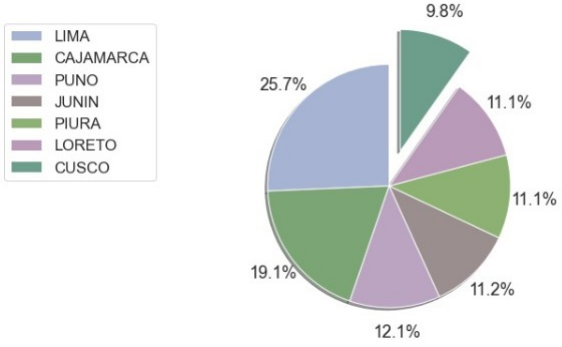

Fig. 1. Top 7 departamentos con más colegios estatales

En la Figura 1, se muestran los 7 departamentos con la mayor cantidad de colegios estatales. Lima destaca 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.

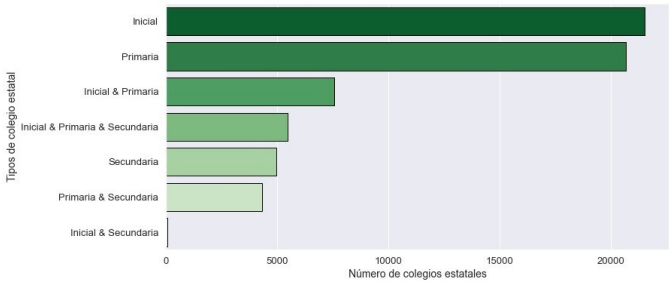

Fig. 2. Tipos de colegios estatales más comunes

En la Figura 2, 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%.

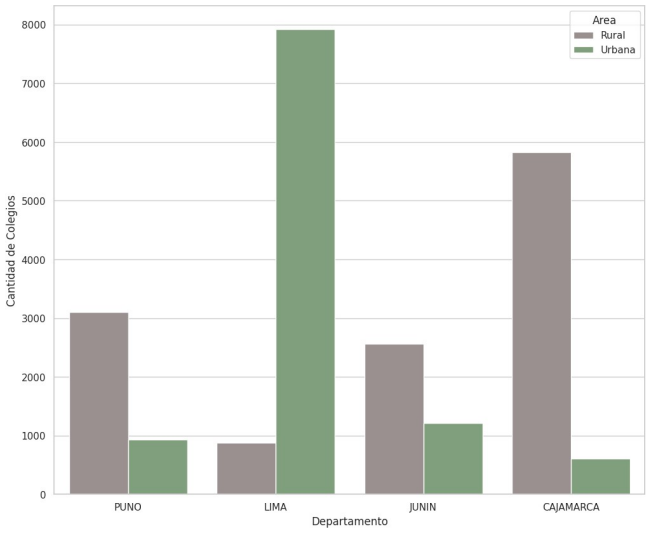

Fig. 3. Proporción de colegios en zona urbana y rural por departamento

En la Figura 3, se presentan datos que muestran los 4 departamentos con la mayor cantidad de colegios. Nuevamente, Lima destaca debido a su considerable población. Algo resaltante es que todos los departamentos excepto Lima cuentan con una mayoría en el área rural.

Continuando con nuestro análisis sobre el ratio entre alumnos y docentes, en la Figura 4 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

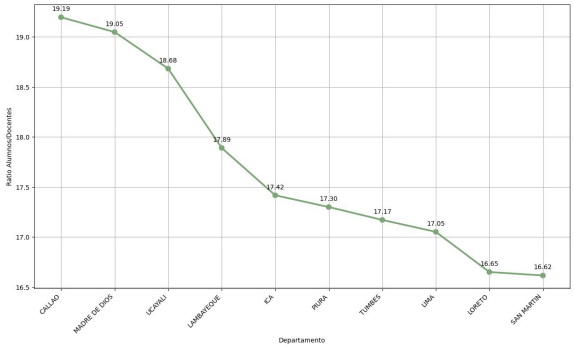

Fig. 4. Ratio de alumno por docente por departamento

negativo en la calidad de la educación. A Callao le siguen departamentos de la costa y la selva. Notablemente, en el top 10, no se encuentra ningún departamento de la sierra.

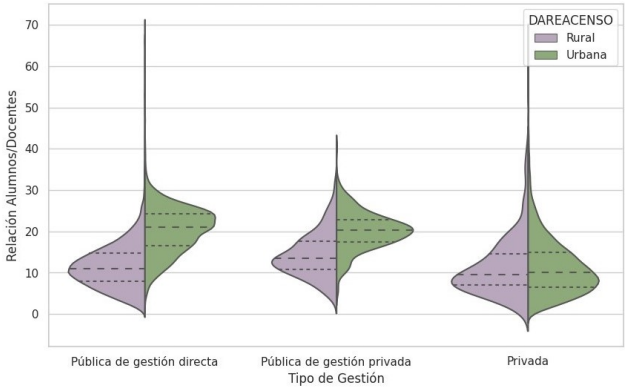

Fig. 5. Distribución del ratio de alumno por docente por tipo de gestión

En la Figura 5, representado en forma de violín, ilustramos 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.

En la Figura 6 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.

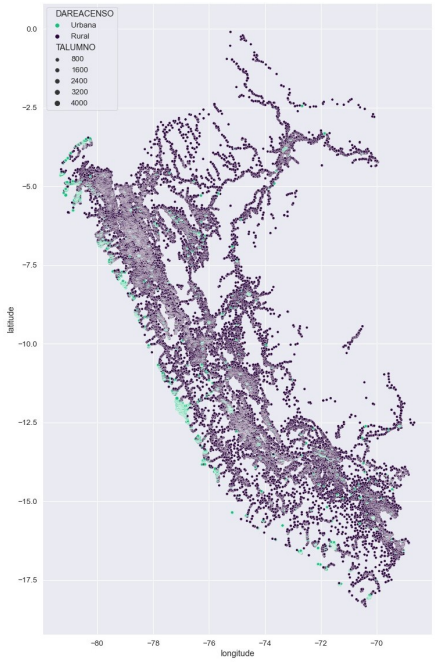

Fig. 6. Distribución de colegios en Perú

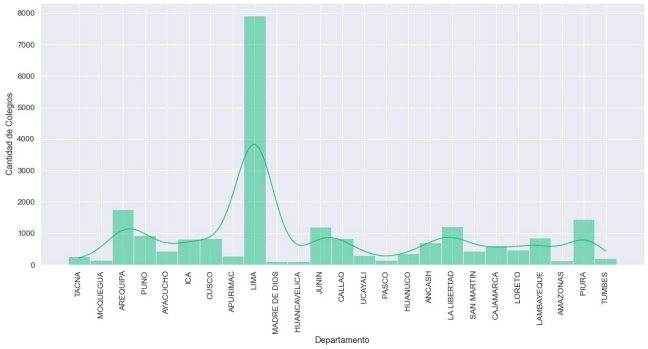

Fig. 7. Distribución de colegios urbanos por departamento

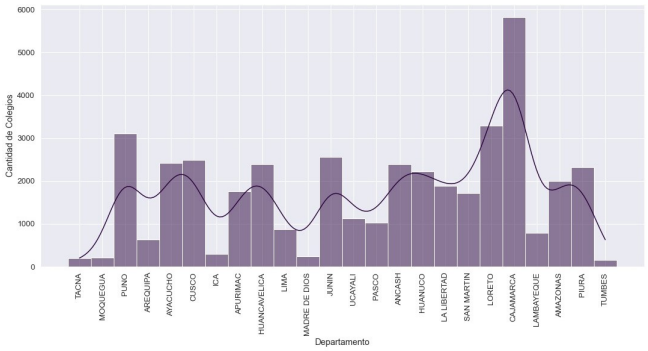

Fig. 8. Distribución de colegios rurales por departamento

VI. Metodología

La metodología que se propone seguir es la de CRISP-DM y se compone de seis etapas: Entendimiento del negocio, entendimiento de la data, preparación de la data, modelamiento, evaluación y despliegue.

1) Entendimiento del negocio:
   - Objetivos del negocio: Garantizar que los estudiantes y las instituciones educativas logren aprendizajes pertinentes y de calidad.
   - Identificación del problema: El acceso a la educación parece ser un privilegio reservado para ciertas zonas geográficas, y cuando se recibe educación, no siempre alcanza los estándares de calidad. Además de esta compleja situación, se añade el desafío del transporte que enfrentan algunos estudiantes para asistir a sus centros educativos, teniendo que desplazarse durante horas para poder acceder a sus clases. Muchos estudiantes se enfrentan a dificultades significativas para desplazarse a la escuela, lo que a menudo involucra caminar largas distancias o utilizar medios de transporte inseguros.
   - ¿Cómo obtener datos?: Los datasets se obtuvieron de la página web del MINEDU y de open source de Google.
   - Formatos de datos: Los datos están estructurados y teniendo en cuenta todos los datasets los campos son de tipo string, float64 y date.
   - Herramientas de software: Se utilizará Python.
2) Entendimiento de los datos:
   - Recopilar datos: Se usan en total 11 dataframes, explicado mas a detalle en secciones previas
   - Descripción de datos: Explicado a detalle en secciones previas
   - Explorar datos: Explicado a detalle en secciones previas
3) Preparación de los datos:
   - Se aplicó selección, integración y limpieza de los datos. Se explicó a detalle en secciones previas.
4) Modelamiento:
   a) Grafico del codo: Se hace el gráfico del codo para 10 repeticiones, 20 repeticiones y 30 repeticiones para poder seleccionar nuestro epsilon óptimo. En este caso, como se ve en la figura 9, el epsilon óptimo es de 0.00115.
   b) Identificar colegios mal ubicados:
      - Para lograr esto primero tenemos que agregarle la columna ”confidence” con valor nulos (NaN) a la latitud y longitud del dataset del MINEDU y una variable etiqueta la cual tendrá el valor de 1, esto para diferenciarlos de los edificios del otro dataset.
      - El segundo paso consiste en agregarle la variable etiqueta al dataset unificado de Open Buildings,

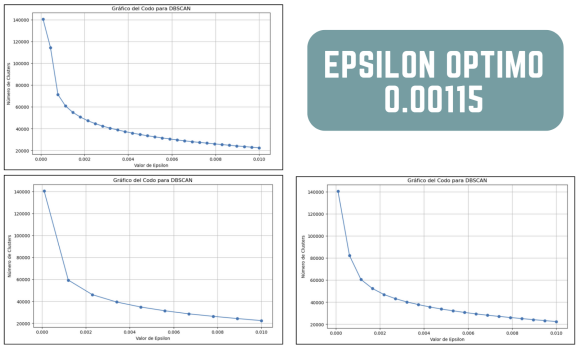

Fig. 9. Gráficos de codo

sin embargo, en este caso el valor de esta columna será de 2.

- El tercer paso es hacer clustering por densidad (DBSCAN) que es un algoritmo de clustering que agrupa puntos de datos basándose en su densidad en lugar de en la distancia. Se inicia seleccionando un punto aleatorio y expande un clúster alrededor de él, incluyendo puntos cercanos que cumplan con ciertos criterios de densidad. Esto permite detectar clústeres de diferentes tamaños y formas sin necesidad de especificar el número de clústeres de antemano.
- El cuarto paso es agregarle el cluster label al dataset al cual se le aplicó el DBSCAN.
- En quinto lugar, eliminamos todos los clusteres los cuales tengan una media de confidence determinada como nula.
- En quinto lugar, filtramos para saber cuales son los clusteres que tienen una media de "confidence" menor a 0.75 (categorizado como no habitable).
- Por último, con los índices substraídos de los colegios en clusteres no habitables (filtramos con etiqueta igual a 1) lo que hacemos es buscar esos índices en el dataset del MINEDU y ya tendríamos todos los datos sobre los colegios mal ubicados.

c) Seleccionar zona para trabajar: En el dataframe de colegios mal ubicados identificamos la zona la cual tiene más colegios clasificados con esa condición y ahí nos vamos a centrar.

d) Delimitar la zona de estudio para los edificios: Se delimita la zona de la región más afectada usando la distancia geodésica y así filtrar los edificios que se encuentran dentro de la región.

e) Encontrar posibles lugares a reubicar: Se vuelve a hacer un DBSCAN, pero solo para la zona que se determinó en el tercer paso. Los resultados de este DBSCAN se filtrarán por los clústeres los cuales tengan un confidence mayor a 0.75 en promedio y no cuenten con un colegio dentro del clúster.

f) Encontrar opciones para reubicar: Se calcula la distancia geodésica entre los colegios mal ubicados y los clústeres aptos para reubicación (los que cumplen con la condición del punto anterior). Por último, se recomiendan las 2 opciones con menor distancia.

5) Evaluación:
   - Evaluación de resultados: Se verificará manualmente en google maps para revisar si es que lo recomendado por el modelo es viable o no.
   - Revisar proceso: Buscar que aplicar técnicas como KNN para poder lidiar con los clusteres que cuentan solo con colegios y no con edificios identificados por open buildings
   - Lista de acciones complementarias:
6) Despliegue:
   - Plan de implementación: Se presentará una interfaz interectiva o un reporte al MINEDU para que ellos luego realicen las evaluaciones pertienentes y se decida si es que es realista reubicar los colegios que están en zonas categorizadas como no habitables.

REFERENCES

[1] E. Hanushek y L. Woessmann, "Education Quality and Economic Growth," Washington, DC: The World Bank, The International Bank for Reconstruction and Development, 2007
[2] P. Glewwe y M. Kremer, "Schools, Teachers, and Education Outcomes in Developing Countries," Handbook on the Economics of Education, Cambridge, Harvard University, 2005.
[3] J. Coleman, E. Q. Campbell, C. J. Hobson, F. McPartland, A. M. Mood, F. D. Weinfeld et al., "Equality of Educational Opportunity," Washington, D.C., U.S. Government Printing Office, 1966.
[4] Banco Mundial, "Primary Education Policy Paper," Washington, D.C., 1990.
[5] M. Fertig y C. M. Schmidt, "The Role of Background Factors for Reading Literacy: Straight National Scores in the PISA 2000 Study," August 2002.
[6] J. Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," NBER Working Paper No. w14442, Oct. 2008
[7] "Relación de instituciones y programas educativos," Ministerio de Educación del Perú, [https://datos.minedu.gob.pe/dataset/instituciones-y-programas-educativos/resource/6f87281b-aa09-44fc-9be1-16a1b255baa4]
[8] "Open Buildings," Google Research, [https://sites.research.google/open-buildings/]