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*—La preocupante prevalencia de infraestructuras escolares inadecuadas y el acceso desigual a una educación de calidad en Perú ha impulsado este proyecto para analizar la ubicación y distribución de las escuelas mediante algoritmos de minería de datos. Al aprovechar un conjunto de datos integral del Ministerio de Educación sobre instituciones educativas, complementado con imágenes satelitales de edificios de Open Buildings, el proyecto tiene como objetivo utilizar distintos algoritmos de tanto data mining como machine learing para identificar escuelas con una infraestructura deficiente en sus alrededores. El objetivo principal es entrenar un algoritmo de machine learning capaz de identificar si es que un colegio se encuentra ubicado en una zona optima para poder ofrecer servicios educativos. En lugar de prescribir políticas específicas, este proyecto busca suministrar análisis clave para informar el proceso de toma de decisiones del gobierno con respecto a las inversiones e intervenciones necesarias para promover un acceso equitativo a una educación de calidad en todo Perú.**

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

I. Introducción

La educación, tal como Hanushek y Woeman [1] han 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 se muestra en el estudio de Coleman et al. [3]. En el estudio de Colemanetal, se empeñan en predecir el éxito académico de los estudiantes estadounidenses, basándose en variables socioeconómicas y escolares de infraestructura. 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] abordan la cobertura del servicio educativo y el ambiente donde se imparte la educación como determinantes que impactan en la calidad del servicio educativo. 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, es decir, una educación que no cumple con lo necesario para que los niños desarrollen todas las competencias que deben durante su época escolar.

En el contexto peruano, el 30% de instituciones educativas en zonas rurales no cuenta con infraestructura adecuada, según datos de la UNESCO [8]. Esta realidad no sólo limita las oportunidades de aprendizaje, sino que también puede afectar psicológicamente a los estudiantes. Los niños y adolescentes son altamente susceptibles a estímulos externos, y la exposición a entornos escolares peligrosos podría derivar en trauma, comportamientos agresivos y dificultades socioemocionales a largo plazo, como plantea la teoría conductista [9].

Ante esta problemática, nuestra motivación es identificar aquellos colegios situados en contextos desfavorables, donde los alumnos estudian en entornos deplorables que atentan contra su bienestar y rendimiento. El propósito es identificar patrones ocultos dentro de la información y desarrollar un modelo de aprendizaje automático que cuente con un accuracy mayor a 0.8 para predecir la habitabilidad de distintas locaciones para la operaración de centros educativos considerando variables que describan el entorno en el cual se encuentra el

colegio.

II. Estado del arte

La optimización de la reubicación de colegios ha sido estudiada intensamente en los últimos años con el fin de poder democratizar la educación, la mayoría de estos estudios se han centrado en cómo variables como la distancia o carreteras afectan a las instituciones educativas. Para este trabajo, se revisó una amplia literatura que muestra enfoques tanto en investigaciones de colegios urbanos como rurales así como estudios en zonas remotas y países de la región. Por ejemplo, revisamos un estudio de reubicación de escuelas rurales en Chile por ingenieros chilenos de la Pontificia Universidad Catolica De Chile y de la Universidad de Chile [13]. Ellos presentaron un modelo de optimización para determinar la ubicación y tamaño óptimos de escuelas rurales en Chile, con el objetivo de apoyar al Ministerio de Educación en la planificación y reestructuración de la infraestructura escolar rural luego del terremoto del 27 de febrero del 2010. El estudio tuvo dos objetivos principales: reducir el número de escuelas multigrado de menor calidad y disminuir las distancias de viaje de los estudiantes, manteniendo costos razonables. El modelo consistió en un programa lineal entero integrado dentro de un sistema de información geográfica. Esto permitió explorar diferentes implicaciones de política educativa, como imponer una distancia máxima de viaje para estudiantes o aumentar la importancia del costo de transporte. El estudio presentó un caso aplicado a todo Chile, proporcionando recomendaciones sobre apertura de nuevas escuelas rurales y expansión, reducción, cierre o mantención de escuelas existentes. También analizó la sensibilidad de las recomendaciones ante cambios en parámetros clave como el costo de transporte. Los resultados mostraron que reubicar escuelas según la optimización podía reducir significativamente las distancias de viaje de estudiantes. El modelo contribuyó a la consolidación de escuelas rurales para lograr un uso más eficiente de los recursos educativos.

Por otro lado en Brasil, un estudio propone determinar ubicaciones óptimas para expandir un sistema de educación superior en el estado de Amazonas, utilizando criterios poblacionales y sociales [11]. El objetivo es evaluar modelos de ubicación de objetivo único para determinar la distribución óptima de instalaciones de educación en ese estado. El estudio sugiere que se necesita una mejor distribución de las IES (Instituciones de Educación Superior). El estudio evalúa y pone a disposición de los tomadores de decisiones tres opciones de optimización: 1) priorizar ciudades con un índice de desarrollo humano de las Naciones Unidas (IDHNU) más bajo; 2) priorizar ciudades con mayor población de estudiantes de secundaria; 3) favorecer ambos criterios por igual. Además, la ubicación debe igualar la distribución de estudiantes entre las cuatro regiones del estado. Con este objetivo, se evaluaron tres modelos de ubicación discretos. El primero de estos modelos, p-center, buscó minimizar la distancia máxima entre los centros de instalaciones y las poblaciones a las que sirven, siendo útil cuando se considera el tiempo o la distancia de viaje, como en la distribución de servicios públicos. El segundo modelo, por su parte, buscó minimizar la suma total de las distancias entre la población y los centros, apuntando a reducir la distancia promedio. En cuanto al tercer modelo, estuvo diseñado para distribuir los centros de manera equitativa entre la población, con el objetivo de minimizar la variabilidad o desequilibrio en la distribución. Los autores concluyen que la metodología desarrollada puede aplicarse para ubicar de manera óptima recursos públicos en general, no solo instituciones educativas. La expansión propuesta favorece el desarrollo social de ciudades con menor IDHNU y permite una distribución más equitativa de estudiantes en el estado.

Otro enfoque ha sido propuesto por investigadores de la Facultad de Ciencias Geográficas de la Universidad Normal de Pekín [15], quienes desarrollaron un método que asigna las plazas escolares combinando la asignación aleatoria con sistemas basados en la proximidad, con el objetivo de lograr una igualdad espacial óptima en las oportunidades educativas. Este concepto se alinea con la noción más amplia de igualdad espacial, que se refiere a la distribución equitativa de recursos, servicios o oportunidades en un área geográfica determinada. En el contexto de la educación, la igualdad espacial de oportunidades educativas significa que todos los estudiantes, independientemente de su ubicación geográfica, tienen acceso a la misma calidad de educación y oportunidades educativas. Al introducir el modelo basado en RES, la investigación ofreció un enfoque innovador que maximizó la similitud entre las distribuciones de probabilidad en los nodos de demanda, considerando restricciones como la distancia máxima de viaje a la escuela y las capacidades. Aplicado a un estudio de caso en el Distrito de Shijingshan en Beijing, este modelo reveló perspectivas prometedoras a través de una resolución heurística. El análisis comparativo con otros modelos, como el modelo capacitado basado en la proximidad y el modelo basado en VAR, destacó la mejora significativa en la igualdad espacial, es decir, se enfoca en la equidad en la distribución de recursos y servicios, mediante la introducción de un mecanismo aleatorio. Sin embargo, esta mejora se produjo a expensas de la eficiencia espacial, es decir, la optimización de la asignación de recursos y servicios para maximizar la eficiencia y minimizar los costos. El estudio concluyó que el modelo basado en RES destacaba como un enfoque óptimo para la igualdad espacial, sin embargo, su implementación futura requeriría consideraciones políticas debido a su posible impacto en la eficiencia espacial y conflictos con la capitalización de la calidad educativa.

Por su parte, investigadores de la provincia de Henan en China [14] plantearon la minimización de costos totales de transporte para los estudiantes, costos de construcción de escuelas, y costos de construcción y mejora de carreteras en una red de tráfico con incertidumbre en los tiempos de viaje indicados por diferentes escenarios. Su estudio planteó que el tiempo de viaje de los estudiantes no debía superar los 45 minutos. Revisaron estudios previos sobre modelos de

ubicación de instalaciones y diseño de redes, señalando que no habían considerado el impacto de la red de tráfico ni las diferencias en costos de viaje por condiciones de caminos en la ubicación de escuelas rurales. Entonces, propusieron un modelo de programación entera mixta y se utilizó un algoritmo híbrido de temple simulado para optimizar la ubicación de escuelas rurales considerando las características de la red de tráfico. Además, se utilizó el enfoque p-robusto estocástico para modelar la incertidumbre en los tiempos de viaje. Esta medida combina los objetivos de la optimización estocástica y la optimización robusta al minimizar el costo esperado mientras se limita el arrepentimiento relativo en cada escenario. En otras palabras, busca encontrar una solución que sea satisfactoria en la mayoría de los escenarios posibles, en lugar de solo en el escenario más probable o en el peor de los casos. El modelo desarrollado determinó eficientemente las ubicaciones óptimas de nuevas escuelas en áreas rurales, la construcción/mejora óptima de enlaces de transferencia, y la asignación óptima de estudiantes a escuelas. Un caso práctico en Guizhou, China confirmó la aplicabilidad del modelo matemático para resolver problemas en la planificación de escuelas rurales. La red de tráfico tuvo una influencia importante en la optimización de ubicaciones de escuelas rurales. Se sugirió incorporar restricciones de capacidad, demanda dinámica y algoritmos para problemas grandes en futuras investigaciones.

Finalmente, en cuanto a la identificación de grupos de edificios, Profesores de la división de Cartografía del Departamento de Ingeniería Geomática de la Universidad Técnica de Yıldız (YTU). [16] compararon cuatro algoritmos de clustering (MST, DBSCAN, CHAMELEON y ASCDT) mediante el índice de Rand ajustado (ARI) y la medida de similitud de Jaccard (JMS). El ARI evalúa la similitud entre dos agrupaciones de datos, con un rango de valores de -1 a 1, donde 1 indicaba una concordancia perfecta, 0 una concordancia aleatoria y -1 una discordancia perfecta. Se obtenía al comparar pares de objetos en ambas agrupaciones y contabilizar cuántos estaban en las mismas o diferentes categorías en ambas agrupaciones, considerando el número de categorías y el tamaño de la muestra. Por otro lado, la JSM también evaluó la similitud entre dos agrupaciones de datos, calculando la proporción de pares de objetos que coincidían en ambas agrupaciones con respecto al número total de pares. Su valor variaba entre 0 y 1, donde 1 indicaba una concordancia perfecta y 0 una discordancia perfecta. A diferencia del ARI, la JSM era una medida más sencilla que no consideraba el número de categorías ni el tamaño de la muestra. Los resultados del estudio indicaron que los modelos de clustering más efectivos para agrupar edificios son DBSCAN y ASCDT.

III. Preguntas de investigación y objetivos

A. Objetivos generales

1) Proporcionar un análisis de datos detallado y presentar recomendaciones fundamentadas al Ministerio de Educación con el propósito de evaluar la idoneidad de la ubicación propuesta para la implementación de un nuevo colegio.

B. Objetivos específicos

1) Utilizar tecnicas de limpieza de datos para poder asegurar efectividad en los algoritmos de data mining y machine learning necesarios para responder a las preguntas de investigación.
2) Identificar escuelas en zonas con infraestructura inadecuada como puede ser la estabilidad de las estructuras mediante análisis de datos.
3) Generar visualizaciones y mapas que resalten disparidades geográficas en el acceso a educación de calidad.
4) Utilizar técnicas de minería de datos como DBSCAN para agrupar y analizar los datos.
5) Utilizar técnicas de Machine Learning para agrupar y comparar modelos y seleccionar el que brinde los mejores resultados.

C. Preguntas de investigación

1) ¿Es posible identificar zonas en Lima donde los colegios presentan deficiencias de infraestructura que limitan el desarrollo de las capacidades establecidas en el currículo nacional?
2) ¿Es factible desarrollar un modelo de aprendizaje automático capaz de predecir la idoneidad de ubicaciones para colegios, con una precisión superior al 80

IV. Marco teórico

En el presente trabajo usaremos distintos algoritmos para cumplir los objetivos propuestos, a continuación se describen algunos de ellos.

- **DBSCAN:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering no paramétrico que se utiliza para encontrar regiones densas de puntos en un espacio de datos. Su funcionamiento se basa en dos parámetros principales: epsilon (ε), que define la distancia máxima entre dos puntos para que se consideren vecinos, y MinSamples, que establece el número mínimo de puntos dentro de ε para formar un "punto central". DBSCAN asigna puntos a diferentes categorías como núcleo, borde o ruido, según su proximidad y densidad. Los puntos centrales se expanden para formar un único grupo, mientras que los puntos de borde se asignan al grupo de un punto central cercano. Los puntos aislados se consideran ruido. Este enfoque permite identificar grupos de formas y tamaños arbitrarios, sin necesidad de especificar el número de clústeres de antemano, lo que lo hace efectivo para conjuntos de datos con estructuras de clústeres complejas y ruido.
- **PCA:** Análisis de componentes principales (PCA por sus siglas en inglés) es un método para reducción de datos redundantes. De acuerdo con Labrín y Urdinez [12], los objetivos de PCA son extraer la información más importante de un dataset, comprimir el tamaño del conjunto de datos manteniendo sólo esta información importante

y analizar la estructura de las observaciones y las variables. Para alcanzar el objetivo de reducción de dimensionalidad, PCA calcula nuevas variables denominadas componentes principales, que son combinaciones lineales de las variables originales. Estas se obtienen mediante el cálculo de eigenvectores y autovalores (eigenvalues) de la matriz de covarianzas. De esta forma, se exige que el primer componente principal capture la mayor cantidad de varianza posible en los datos. Luego, cada componente sucesivo se calcula bajo la restricción de ser ortogonal, es decir, que no esté correlacionado a los componentes anteriores, y capturando la máxima varianza restante. De esta forma, los primeros componentes principales sintetizan la variabilidad más importante de todas las variables originales. Los últimos componentes modelan solo ruido y variabilidad residual sin importancia, por lo cual pueden eliminarse sin gran pérdida de información.

- **KNN:** KNN (K-Nearest Neighbors) es un algoritmo de aprendizaje supervisado que se utiliza para clasificación y regresión. Funciona asignando una etiqueta a un punto desconocido basándose en las etiquetas de los puntos vecinos más cercanos en el espacio de características. La predicción se realiza calculando la distancia entre el punto de consulta y todos los demás puntos en el conjunto de datos de entrenamiento. Los k puntos más cercanos se seleccionan y se utiliza la mayoría de votos (en el caso de clasificación) o el promedio (en el caso de regresión) de los valores conocidos para predecir la etiqueta o el valor del punto desconocido. KNN es sencillo de implementar y entender, aunque su rendimiento puede ser sensible a la elección del parámetro k y a la dimensionalidad de los datos, siendo más efectivo en conjuntos de datos con distribuciones bien definidas y sin ruido.
- **Random Forest:** Random Forest (Bosques Aleatorios) es un algoritmo de aprendizaje supervisado que se utiliza tanto para tareas de clasificación como de regresión. Funciona construyendo múltiples árboles de decisión durante el proceso de entrenamiento y combinando sus predicciones para obtener una predicción final más precisa y robusta. Cada árbol en el bosque se entrena con una muestra aleatoria del conjunto de datos de entrenamiento y utiliza una selección aleatoria de características para dividir los nodos, lo que ayuda a reducir el sobreajuste y mejorar la generalización. Durante la predicción, las predicciones individuales de cada árbol se combinan mediante votación en el caso de clasificación o promedio en el caso de regresión, para obtener una predicción final. Random Forest es conocido por su capacidad para manejar conjuntos de datos grandes con alta dimensionalidad, su resistencia al sobreajuste y su capacidad para manejar datos faltantes y variables categóricas sin necesidad de preprocesamiento adicional. En el contexto de clasificación, los datos etiquetados son esenciales para entrenar el modelo. A partir de estos datos de entrada y sus correspondientes etiquetas, el algoritmo crea un modelo capaz de clasificar nuevas entradas en categorías predefinidas. La importancia de Random Forest en clasificación radica en su capacidad para manejar eficazmente problemas de clasificación con múltiples clases, así como su habilidad para evaluar la importancia de las características en la toma de decisiones. Esto lo convierte en una opción popular y confiable para una amplia gama de aplicaciones de aprendizaje automático.
- **t-SNE:**t-SNE(t-Distributed Stochastic Neighbor Embedding): Es una técnica no supervisada de reducción de dimensionalidad utilizada para la visualización de datos de alta dimensión. Funciona transformando los datos de entrada en un espacio de baja dimensión (2D o 3D), de forma que las distancias entre puntos cercanos se preserven mientras que las distancias entre puntos distantes se hacen más grandes. Esto permite revelar la estructura de clusters o grupos dentro de los datos.
- **Apriori:**Es un algoritmo de reglas de asociación (association rule learning) frecuentemente utilizado en minería de datos. Identifica relaciones interesantes entre variables en grandes bases de datos, encontrando conjuntos de elementos que ocurren juntos frecuentemente y generando reglas que predicen la ocurrencia de un elemento basado en la ocurrencia de otros. Por ejemplo, puede identificar que los clientes que compran producto A y B también tienden a comprar el producto C.
- **Matriz de confusón:** La matriz de confusión es una herramienta esencial en la evaluación de modelos de clasificación. Esta matriz organiza las predicciones del modelo en cuatro categorías: Verdaderos Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) y Verdaderos Negativos (VN). Los Verdaderos Positivos representan instancias correctamente clasificadas como positivas, mientras que los Falsos Positivos indican instancias incorrectamente clasificadas como positivas. Por otro lado, los Falsos Negativos son instancias erróneamente clasificadas como negativas, y los Verdaderos Negativos son instancias correctamente clasificadas como negativas. Estos elementos proporcionan una visión detallada del rendimiento del modelo, permitiendo calcular métricas cruciales como precisión, sensibilidad y especificidad.

V. METODOLOGIA

A. 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, recuperado del Repositorio de Datos Abiertos del Ministerio de Educación [7]. Y la segunda es un dataset de edificaciones de Africa, Asia, América Latina y el Caribe, resultado del proyecto Open Buildings de Google Research [10]. 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 implementó un filtro adicional con el fin de seleccionar exclusivamente aquellas instituciones que proporcionan educación escolarizada, excluyendo otras modalidades educativas. El término "modalidad escolarizada" se refiere a los colegios que imparten clases de manera presencial, siguiendo el esquema clásico de educación establecido por el ministerio de educación, abarcando desde 1ro hasta 6to de primaria y de 1ro a 5to de secundaria en un periodo de 11 años.
3) **Filtrado por Niveles Educativos:** Se decidió conservar en el conjunto de datos únicamente las instituciones correspondientes a los niveles educativos de Inicial, Primaria y Secundaria, excluyendo las demás categorías educativas. Esta elección se basa en que estos niveles comprenden el 87% de la población estudiantil en el Perú y representan los niveles más tradicionales de educación.
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 unicos útiles para la exploración y explotación de la data.

- **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_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.

Luego de analizar los atributos, se decidió utilizar solo 3 atributos, ya que son los mas apropiados para poder complementar la Base de Datos de colegios. Los atributos son:

- **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 0.775 o superior para considerarla habitable.

Estos procedimientos aseguran la relevancia y pertinencia de los datos para el análisis y la extracción de conocimiento. Como consecuencia, ahora contamos con conjuntos de datos que incluyen información significativa sobre la idoneidad de la ubicación de los colegios.

B. 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.

En la Figura 1, se muestra un diagrama de pastel que muestra los 7 departamentos principales en Perú con la mayor cantidad de colegios estatales. El porcentaje de colegios estatales en cada departamento se representa como una porción del pastel. Donde, Lima tiene el 25.9% de las escuelas estatales, lo que es la porción más grande del pastel.

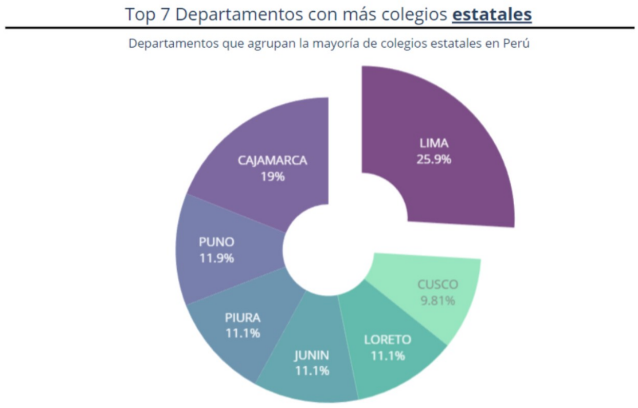

Fig. 1. Top 7 departamentos con más colegios estatales

La relevancia de este gráfico radica en su capacidad para mostrar cuáles departamentos en Perú tienen la mayor cantidad de colegios estatales y donde el gobierno puede focalizar sus esfuerzos de mejora.

La figura 2 presenta un diagrama de violín que muestra la distribución de la relación entre estudiantes y profesores en diferentes tipos de gestión: públicas de gestión directa, públicas de gestión privada y privadas. Se muestra la relación de estudiantes con profesores. Las flechas indican las relaciones máximas y mínimas de estudiantes a profesores para cada tipo de gestión. Este gráfico es útil para entender cómo se comparan estos tipos de escuelas en términos de tamaño de clase, lo que puede tener implicaciones para la calidad de la educación. En el gráfico, es notable un mayor ratio en áreas urbanas en los tres tipos de gestión. Además, es destacable y preocupante los valores máximos en pública de gestión directa y privada. Un ratio de alrededor 70 alumnos por docente puede afectar la calidad de la educación.

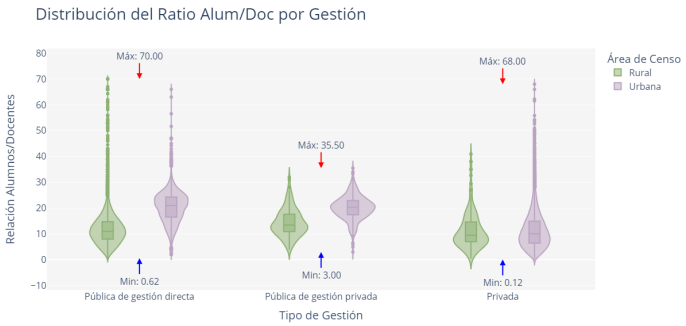

Fig. 2. Distribución del Ratio Alumno/Docente por Gestión

El gráfico 3 presenta un diagrama de barras que ilustra la distribución del número de escuelas estatales en distintas categorías. En el eje X, se representa el número de escuelas estatales, mientras que en el eje Y se muestran las categorías de escuelas, que incluyen ”Inicial & Secundaria”, ”Primaria & Secundaria”, ”Inicial & Primaria”, ”Primaria” e ”Inicial”. Cada barra refleja la cantidad de escuelas en una categoría específica, y la longitud de estas barras proporciona una visualización directa de la cantidad de instituciones educativas en cada categoría.

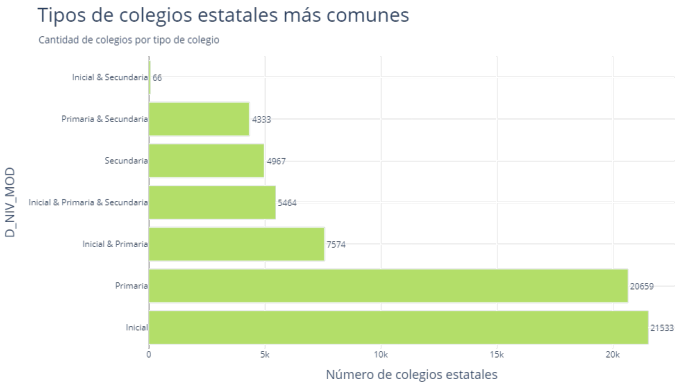

Fig. 3. Tipos de colegios estatales más comunes

Este gráfico facilita la comprensión detallada de la composición de la oferta educativa del país, destacando que la categoría ”Inicial” tiene la mayor cantidad de colegios, mientras que la categorías que contienen a ”Secundaria” representa un porcentaje menor incluso a Primaria.

El gráfico 4 se presenta como un mapa de dispersión que ilustra la distribución de las escuelas en Perú. Cada punto en el mapa representa una institución educativa, y el tamaño de los puntos guarda proporción con el número de estudiantes en cada escuela. De esta manera, las escuelas con mayor cantidad de estudiantes se representan con puntos más grandes, mientras que los colores varían según la ubicación: morado para áreas urbanas y verde para áreas rurales.

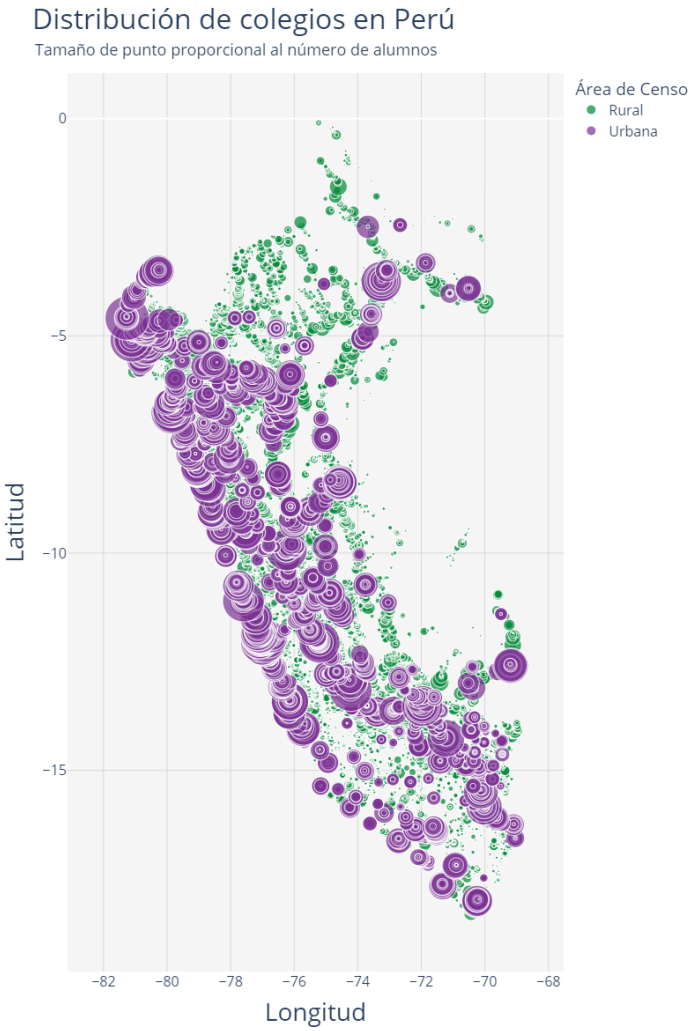

Fig. 4. Distribución de colegios en Perú

La importancia de este gráfico radica en su capacidad para visualizar la distribución geográfica de las escuelas en Perú y la cantidad de estudiantes en cada una. Esta información resulta invaluable para comprender la disposición de la educación en el país, permitiendo identificar áreas que podrían requerir intervenciones educativas adicionales.

El gráfico 5 representa la distribución de escuelas en los 10 departamentos de Perú con la mayor cantidad de colegios unidocentes, clasificados según áreas rurales y urbanas. Las barras se han coloreado en verde para representar los colegios en áreas rurales y en azul para los ubicados en áreas urbanas. La longitud de cada barra refleja el número de escuelas en cada departamento. Loreto y Cajamarca son los departamentos con la mayor cantidad de colegios unidocentes.

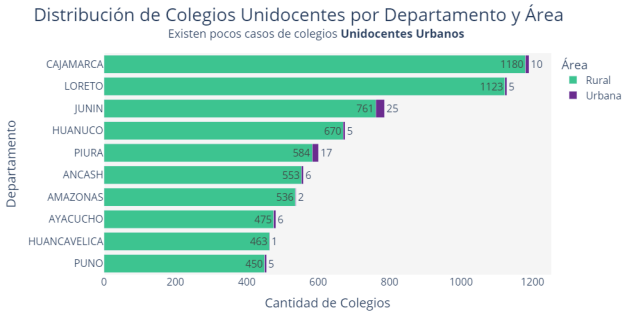

Fig. 5. Distribución de Colegios Unidocentes por Departamento y Área

El gráfico destaca las regiones donde los colegios cuentan únicamente con un profesor en su plantel, lo cual representa un desafío significativo para la calidad de la enseñanza en estas instituciones.

C. Data Mining y Machine learning

El objetivo del trabajo es predecir si una ubicación es adecuada para un colegio. Para lograrlo, es necesario identificar tanto la variable dependiente como las independientes. La variable dependiente es el nivel de confianza del colegio. Para las variables independientes, es importante identificar los factores que podrían influir en el estado de la infraestructura del colegio y los edificios cercanos. Por lo tanto, se divide la metodología en dos etapas: primero, se calcula el nivel de confianza (el valor de confidence) de cada colegio y luego se identifican las variables que predicen este nivel de confianza.

1) Obtención del valor de confidence de los colegios:
   - El primer paso consiste en preparar los conjuntos de datos para identificar los colegios junto con los edificios. La base de datos de colegios se reduce a las coordenadas de latitud y longitud, mientras que en la base de datos de Open Buildings se conservan los campos de latitud, longitud y confianza (”confidence”).
   - Con las bases de datos debidamente preparadas, se procede a examinar coincidencias exactas entre las coordenadas de latitud y longitud de los edificios y colegios. Para ello, se concatenan los conjuntos de datos y se filtran para que las coordenadas de latitud y longitud de los colegios coincidan con las de los edificios. Sín embargo, no se encuentran coincidencias exactas debido a la extrema precisión de las mediciones de latitud y longitud, donde un pequeño desplazamiento puede afectar la correspondencia en las bases de datos. Por este motivo, se recurre a la agrupación por densidad (cluster por densidad) para identificar puntos cercanos.
   - Como se explicó anteriormente, el algoritmo DBSCAN requiere la configuración de dos parámetros para funcionar correctamente. El valor de min_samples se establece en 5 para formar

clusters que representen edificios cercanos a los colegios. En cuanto al parámetro epsilon, se determina a través de la gráfica del codo, visible en la Figura 6 con un valor de 0.0005.

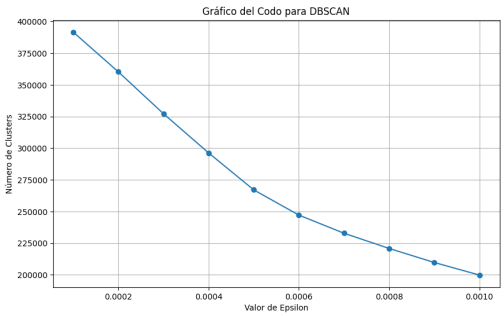

Fig. 6. Gráficos de codo

- Con los parámetros definidos, se aplica la técnica al conjunto de datos combinado de colegios y edificios en Perú. Se asignan los valores promedio de confianza a cada colegio. Sin embargo, como se observa en la figura 7, algunos colegios aún tienen un valor nulo en la columna de confianza. Esto se debe a que los clústeres están formados por un solo colegio.

```
latitude            0
longitude           0
confidence       5293
Etiqueta            0
cluster_label       0
dtype: int64
```

Fig. 7. Colegios con confidence nulo

- Para evaluar estos casos, se identifican los cinco vecinos más cercanos y se asigna el promedio de sus niveles de confianza. Se tiene en cuenta la distancia geodésica, limitada a no superar 1 km. Como se visualiza en la figura 12, todos los colegios tienen ahora asignado un valor de confianza. Finalmente, se etiquetan como 'Bien ubicados' aquellos colegios con un valor de confianza superior a 0.775, mientras que aquellos con un valor igual o inferior a 0.775 se consideran 'Reubicables'.

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

Fig. 8. Colegios con confidence nulo despues de tratar los nulos

2) Obtención de variables predictoras:
En la obra de Cubillos-González [16], la habitabilidad de un edificio se conceptualiza como la capacidad para garantizar condiciones mínimas de confort y salubridad a sus residentes. Garay [17] argumenta que el volumen de habitantes en un edificio, el comportamiento de estos residentes, y los aspectos técnicos inherentes a la construcción pueden influir significativamente en su habitabilidad.
Entre los aspectos técnicos, se destacan el tipo de suelo, los materiales utilizados, así como el acceso a servicios de alcantarillado y electricidad, entre otros. Por otro lado, el comportamiento de los residentes desempeña un papel fundamental, siendo la proximidad a comercios y centros médicos un factor que fomenta la actividad alrededor del edificio, asociándose, según Thompson [19], con mayor vigilancia natural y menos incidencias delictivas, mejorando así la seguridad y habitabilidad.
De manera similar, la cercanía a parques y áreas verdes incentiva actividades recreativas fuera del edificio, impulsando la cohesión social, el sentido de comunidad y el cuidado del entorno, elementos positivamente vinculados con la habitabilidad según Willams et al. [20].
Dada la complejidad de obtener información exacta sobre crímenes, materiales de construcción y mantenimiento de colegios, se optó por seleccionar las siguientes variables: tiendas cercanas, hospitales cercanos, comisarías cercanas, parques cercanos y la cantidad de personas constantes en el colegio.
Las primeras cuatro se obtendrán mediante la librería osmnx para almacenar el grafo de Lima provincia y así identificar comisarías, tiendas, hospitales y parques.
Respecto a la última variable, se obtiene sumando el número de profesores y alumnos en el colegio.

a) A través de la implementación de la librería OSMX, se procederá a almacenar el grafo correspondiente a la provincia de Lima con el objetivo de identificar ubicaciones clave, tales como comisarías, tiendas, hospitales y parques.
b) Cada institución educativa será asociada con el número de comisarías, tiendas, hospitales y parques que se encuentren a una distancia inferior a 1 kilómetro. Se crea además una nueva variable, densidad poblacional, se obtendrá sumando la cantidad de profesores y alumnos en cada colegio.
c) El primer paso consiste en seleccionar las cinco variables predictoras más relevantes, junto con la variable "confidence", que actuará como la etiqueta o variable dependiente para el entrenamiento de los modelos. Aquellos colegios con un valor de "confidence" menor a 0.775 serán etiquetados como "Reubicables", indicando que se encuentran en áreas consideradas inadecuadas. Por otro lado, aquellos con un valor de "confidence" igual o superior a 0.775 serán clasificados como "Bien ubicados".
d) Adicionalmente, se seleccionan las variables predictoras que se consiguieron de OSMX para

aplicarles PCA de tres componentes y estos nuevos datos agregarlos al dataframe principal para darle robustez a la predicción

Posteriormente, se procederá a dividir el conjunto de datos en subconjuntos de entrenamiento y prueba. Una vez que los datos estén separados y etiquetados, se llevará a cabo el entrenamiento de modelos de aprendizaje supervisado.

Durante la fase de entrenamiento, estos modelos aprenderán a predecir la variable ”confidence” basándose en las cinco variables predictoras previamente seleccionadas. Una vez completado el entrenamiento, los modelos estarán listos para realizar predicciones sobre nuevos datos cuya habitabilidad sea desconocida. Se realizará una comparación de métricas de rendimiento, como precisión, recall y el F1-Score, con el fin de determinar cuál de los modelos presenta una mejor capacidad de generalización.

VI. Resultados

Los resultados son presentados luego de tener lists la base de datos de los colegios con su respectivo confidence asignado.

1) **Exploración:**
   a) **T-SNE:** En la Figura 12, se puede visualizar cómo no hay una diferencia clara entre las clases de la variable ’confidence’ para los colegios ’Reubicables’ y ’Bien ubicados’. Como se observa en la Figura 12, a pesar de variar el parámetro ’perplexity’, las clases no llegan a ser distinguibles. Surge la incógnita sobre si existe una relación clara entre las clases y las variables predictoras.

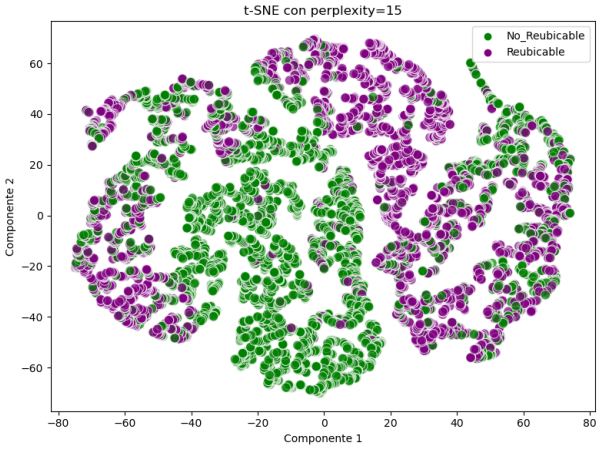

Fig. 9. Matriz de confusión de Random Forest Classifier

   b) **Algoritmo de Asociación:** Se utilizará el algoritmo Apriori para identificar patrones simples, precisos y útiles. Antes de utilizar el algoritmo, es necesario preprocesar los datos. Primero, se deben crear etiquetas para todas las variables predictoras numéricas. En este caso, las etiquetas se generaron considerando el cuartil al que pertenece cada valor, utilizando el siguiente criterio:

Valor perteneciente a	Label
Cuartil 3	Alto
Cuartil 2	Medio
Cuartil 1	Bajo

TABLE I
CRITERIO DE ETIQUETADO PARA VARIABLES PREDICTORAS

Con los datos ya preprocesados, se aplicó el algoritmo Apriori y se descubrieron los siguientes patrones:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
250	(confidence_Reubicable, densidad_poblacional_e...	(hospitales_cercanos_escala_Medio, parques_cer...	0.039229	0.249756	0.027502	0.701068	2.807014
241	(hospitales_cercanos_escala_Alto, confidence_R...	(tiendas_cercanas_escala_Alto)	0.027223	0.338406	0.022197	0.815385	2.409488
240	(hospitales_cercanos_escala_Alto, confidence_R...	(comisarias_cercanas_escala_Alto)	0.025548	0.363674	0.022197	0.868852	2.389094

Fig. 10. Asociación de colegios ’Reubicables’ como antecedente

Como se observa en la Figura 10, al filtrar los colegios ’Reubicables’ como antecedente, se encontró que la probabilidad de que las comisarías cercanas tengan un nivel alto se incrementa en un 138%, contradiciendo la creencia popular. Además, en la Figura 11, los colegios ’Bien ubicados’ se asocian con un nivel medio de cercanía a comisarías.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
277	(hospitales_cercanos_escala_Medio, densidad_po...	(comisarias_cercanas_escala_Medio)	0.039927	0.636326	0.037275	0.933566	1.467121
272	(hospitales_cercanos_escala_Medio, tiendas_cer...	(comisarias_cercanas_escala_Medio)	0.028899	0.636326	0.026944	0.932367	1.465236
206	(tiendas_cercanas_escala_Bajo, hospitales_cerc...	(comisarias_cercanas_escala_Medio)	0.104007	0.636326	0.096608	0.928859	1.459723

Fig. 11. Asociación de colegios ’Bien ubicados’ como antecedente

En resumen, al tener como antecedente colegios ’Bien ubicados’, se tienen 21 atributos clasificados como bajo, 77 como medio y 3 como alto. Por otro lado, si se tiene como antecedente un colegio ’Reubicable’, se tienen 12 atributos clasificados como bajo, 63 como medio y 10 como alto.

2) **Explotación:** La implementación del Random Forest Classifier revela un rendimiento robusto en la clasificación tanto en el conjunto de entrenamiento, compuesto por 5730 instancias, como en el conjunto de prueba con 1433 registros. Detallando las métricas, se logra una precisión del 87.34%, un recall del 88%, y un F1 score del 87%, destacando su capacidad para predecir instancias positivas y negativas. En el conjunto de prueba, se observa una precisión del 81.79%, un recall del 82%, y un F1 score del 82%. Aunque demuestra habilidad para realizar predicciones, se nota un ligero sobreajuste, posiblemente atribuido a la naturaleza de las variables

utilizadas y la necesidad de incorporar otras para una mejora adicional.

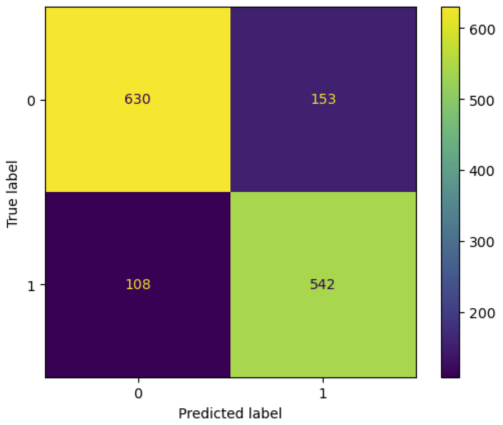

Fig. 12. Matriz de confusion de Random Forest Classifier

La importancia de este modelo reside en su capacidad para llevar a cabo una clasificación precisa. Esto se vuelve importante, cuando se considera la seguridad de estudiantes y docentes al evaluar la idoneidad de la ubicación de una escuela. La información incorrecta sobre si un colegio está en una buena ubicación puede generar riesgos significativos para quienes asisten a la institución. Por lo tanto, la precisión del modelo es fundamental para garantizar un entorno seguro para todos los involucrados en el ámbito educativo.

VII. Conclusiones

1) Es posible utilizar técnicas de minería de datos como DBSCAN y machine learning para identificar escuelas con infraestructura deficiente en sus alrededores. El paper desarrolla un modelo de aprendizaje automático capaz de predecir la idoneidad de ubicaciones para colegios con una precisión superior al 80
2) Entre los factores que predicen la habitabilidad de ubicaciones para colegios se encuentran: cercanía a comisarías, tiendas, hospitales y parques, así como la densidad poblacional del colegio. Sin embargo, se requiere una investigación más profunda sobre las variables que expliquen mejor el fenomeno de la habitabilidad de un colegio
3) Los resultados del algoritmo de asociación contradicen la creencia popular de que mayor cercanía a comisarías se asocia con mejor ubicación de los colegios. Esto resalta la complejidad del problema y la necesidad de un enfoque integral, considerando múltiples aristas.

VIII. Limitaciones

1) El trabajo requiere de un gran costos computacional debido a que hay muchos edificios y colegios en Lima, por lo que resultó inviable poder recolectar todas las variables predictoras para todo el país.
2) No se cuenta con información sobre los delitos a nivel de latitud y longitud para poder tener una mejor medición sobre el nivel de seguridad que hay en una zona y no recurrir a la cercania a comisarias u otros lugares que puedan explicar la seguridad de una zona.

IX. Futuros trabajos

1) Un trabajo futuro puede ser explicar la calidad educativa buscando factores adicionales a la infraestructura del colegio.
2) Como complemento al trabajo presentado se recomienda hacer un filtro hirbido con la final de usar el filtro colaborativo para asignar un valor de confidence y crear un perfil de colegio para luego aplicar filtro basado en contenido y buscar clusteres los cuales se ajusten al perfil del colegio que se requiere

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] UNESCO, *Informe de Seguimiento de la Educación en el Mundo 2019: Migración, Desplazamiento y Educación - Construyendo Puentes, no Muros*. París: UNESCO, 2019.

[9] B. F. Skinner, *Beyond Freedom and Dignity*. Londres: Penguin, 1971.

[10] "Open Buildings," Google Research, [https://sites.research.google/open-buildings/]

[11] C. M. Xavier, M. G. Fernandes Costa, and C. F. F. C. Filho, "Combining Facility-Location Approaches for Public Schools Expansion," in *IEEE Access*, vol. 8, pp. 24229-24241, 2020, doi: 10.1109/ACCESS.2020.2970385.

[12] C. Labrín and F. Urdinez, "Principal component analysis," in R for Political Data Science, Chapman and Hall/CRC, pp. 375-393, 2020.

[13] F. Araya, R. Dell, P. Donoso, V. Marianov, F. Martínez, and A. Weintraub, "Optimizing location and size of rural schools in Chile," *Intl. Trans. in Op. Res.*, 19: 695-710, 2012.

[14] Y. Chen, "Optimizing Locations of Primary Schools in Rural Areas of China," *Complexity*, vol. 2021, Article ID 7573700, 13 pages, 2021.

[15] T. Dai, C. Liao, and S. Zhao, "Optimizing the spatial assignment of schools through a random mechanism towards equal educational opportunity: A resemblance approach," *Computers, Environment and Urban Systems*, vol. 76, pp. 24–30, 2019.

[16] S. Cetinkaya, M. Basaraner, and D. Burghardt, "Proximity-based grouping of buildings in urban blocks: a comparison of four algorithms," *Geocarto International*, vol. 30, no. 6, pp. 618-632, 2015.

[17] R. M. Garay, R. Tapia, M. Castillo, O. Fernández, y J. Vergara, "Habitabilidad de edificaciones y ranking de discriminación basado en seguridad y sustentabilidad frente a eventuales desastres. Estudio de caso: Viviendas de madera," *Revista de Estudios Latinoamericanos sobre Reducción del Riesgo de Desastres (REDER)*, vol. 2, no. 2, pp. 28-45, 2018.

[18] R. A. Cubillos-González y C. M. Rodríguez-Álvarez, "Evaluación del factor de habitabilidad en las edificaciones sostenibles," Grupo de investigación sostenibilidad, medio ambiente y tecnología en arquitectura, SOMET, Universidad Católica de Colombia. Fecha de recepción: 24/04/2013. Fecha de aceptación: 15/06/2013.

[19] S. Thompson, "Natural Surveillance," en *Environmental Criminology and Crime Analysis*, pp. 71-84, Routledge, 2020.

[20] A. Williams, V. Ceccato, B. Söderberg, y M. Wilhelmsson, "The impact of neighbourhood green qualities on physical activity and mental health of citizens at risk of inactivity," *Health & Place*, vol. 58, p. 102017, 2019.