

UNIVERSIDAD PRIVADA “FRANZ TAMAYO”

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA DE SISTEMAS



**“ANÁLISIS DE DATOS CENSALES PARA DETERMINAR LA UBICACIÓN
OPTIMA Y NECESARIA DE CENTROS EDUCATIVOS
EN BOLIVIA”**

ESTUDIANTE: Gabriela Micaela
Durán Villafán

ASIGNATURA: Big Data

DOCENTE: Ing. Enrique Laurel

LA PAZ – BOLIVIA

2025

1. Introducción

La planificación urbana y la asignación eficiente de recursos públicos son pilares fundamentales para el desarrollo social y económico de cualquier país. En este contexto, la educación se erige como un derecho básico y un motor de movilidad social, cuya accesibilidad y calidad dependen en gran medida de una distribución territorial equitativa de los centros educativos. Sin embargo, la dinámica demográfica, caracterizada por constantes flujos migratorios y cambios en la densidad poblacional, genera desfases entre la oferta educativa existente y las necesidades reales de la población.

Tradicionalmente, las decisiones sobre la localización de nuevas escuelas o la ampliación de las existentes se han basado en métodos cualitativos, estudios de demanda parciales o criterios políticos, lo que a menudo resulta en una asignación inefficiente de los recursos. Esto puede traducirse en "desiertos educativos" en algunas zonas, donde los niños y jóvenes deben recorrer grandes distancias para acceder a la educación, y en una subutilización de la infraestructura en otras.

La era del Big Data ofrece una oportunidad sin precedentes para revolucionar este campo. Los censos de población y vivienda, junto con otras fuentes de datos administrativos, contienen un vasto y rico ecosistema de información sobre la composición demográfica, socioeconómica y geográfica de un territorio. El análisis de estos macrodatos permite pasar de la intuición a la evidencia, de la planificación estática a la dinámica.

Este proyecto propone el uso de técnicas avanzadas de análisis de datos y ciencia de datos para procesar información censal, con el fin de identificar de manera precisa, objetiva y cuantificable las áreas geográficas con mayor necesidad de nuevos centros educativos. El objetivo final es proporcionar a los planificadores y autoridades una herramienta de soporte decisional basada en datos que optimice la inversión pública y garantice el derecho a la educación.

2. Justificación

Importancia desde la Perspectiva de Big Data

Este proyecto es inherentemente un problema de Big Data. Los censos nacionales manejan volúmenes masivos de información (Volume) que abarcan a millones de individuos y hogares. La variedad de datos es enorme, incluyendo números (edad, ingresos), texto (direcciones, nivel educativo), y ubicaciones geográficas (Variety). La velocidad de análisis, aunque no es en tiempo real, requiere procesar estos datos de manera eficiente para generar insights útiles en plazos razonables (Velocity). Además, la veracidad de los datos oficiales los convierte en una fuente confiable (Veracity), y el potencial valor de la información extraída para la toma de decisiones es incalculable (Value). Sin las herramientas del Big Data, explotar todo el potencial de estos conjuntos de datos sería prácticamente imposible.

Justificación

Técnica:

Se emplearán tecnologías y metodologías propias del ecosistema Big Data. Esto incluye el uso de frameworks de procesamiento distribuido como Apache Spark para manejar el volumen de datos, sistemas de almacenamiento en clústeres (HDFS) o soluciones en la nube (AWS S3, Google BigQuery). El análisis espacial se realizará con librerías geoespaciales (GeoPandas en Python) y Sistemas de Información Geográfica (GIS). Se aplicarán algoritmos de clustering (como DBSCAN o K-Means) para identificar aglomeraciones de población en edad escolar y modelos de optimización de localización (Análisis P-Mediano) para determinar la ubicación óptima que minimice la distancia promedio para los estudiantes.

Justificación

Social:

Una distribución óptima de los centros educativos reduce las brechas de inequidad, asegurando que todos los niños y jóvenes, independientemente de su lugar de residencia, tengan acceso a una educación de calidad cerca de sus hogares. Esto disminuye los tiempos de desplazamiento, reduce los costos de transporte para las familias, aumenta la seguridad de los estudiantes y fomenta una mayor integración

comunitaria. El proyecto contribuye directamente a la consecución de los Objetivos de Desarrollo Sostenible (ODS), específicamente al ODS 4: "Educación de Calidad".

Justificación

Económica:

La construcción de un centro educativo representa una inversión pública millonaria. Una mala localización puede derivar en un alto costo de oportunidad y en el subaprovechamiento de la infraestructura. Este análisis permite priorizar la inversión en las áreas de mayor impacto demográfico, asegurando que cada peso invertido se dirija a satisfacer una necesidad real y crítica. Además, optimiza los costos operativos futuros (transporte escolar, mantenimiento) y previene la construcción de infraestructura subutilizada que generaría gastos innecesarios a largo plazo.

3. Planteamiento del Problema

Descripción del Problema

Actualmente, la planificación de la infraestructura educativa en muchas regiones no responde de manera ágil y precisa a los cambios demográficos. Esto genera una distribución geográfica desequilibrada de la oferta educativa, manifestándose en dos problemas principales:

- Sobresaturación de escuelas en zonas donde la demanda ha decrecido, llevando a una subutilización de recursos.
- Carencia de establecimientos en zonas de crecimiento poblacional reciente, forzando a los estudiantes a largos y costosos desplazamientos o limitando su acceso a la educación.
- La falta de una herramienta basada en datos objetivos y de alcance nacional/municipal perpetúa esta ineficiencia, afectando la calidad de vida y el desarrollo educativo de la población.

4. Objetivos de Investigación

Objetivo

General

Desarrollar un modelo de análisis basado en Big Data que, mediante el

procesamiento de datos censales y geoespaciales, identifique las ubicaciones más óptimas y prioritarias para la construcción de nuevos centros educativos en una región determinada.

Objetivos Específicos

- Recopilar, limpiar e integrar los conjuntos de datos censales y la ubicación geográfica de los centros educativos existentes.
- Geocodificar las direcciones de la población objetivo y los centros existentes para su representación espacial.
- Realizar un análisis de densidad espacial para identificar "departamentos" con alta concentración de población en edad escolar y baja oferta educativa.
- Aplicar algoritmos de optimización para localización de servicios que indiquen los departamentos ideales para nuevos centros.
- Visualizar los resultados mediante mapas interactivos y dashboards que presenten de forma clara las recomendaciones a los tomadores de decisiones.

5. Fuentes de Datos

- **Censo Nacional de Población y Vivienda:** Principal fuente de datos demográficos.
- **Encuestas de Hogares:** Para complementar y actualizar información entre periodos censales.
- **Registro Nacional de Instituciones Educativas:** Para la geolocalización y capacidad de las escuelas existentes.

6. Cronograma Estimado (4 Meses)

Mes	Fase y Actividades Principales
	Fase 1: Adquisición y Preprocesamiento
Mes 1	<ul style="list-style-type: none"> - Recopilación de datos censales y geoespaciales. - Limpieza, integración y transformación de datos (ETL). - Geocodificación de direcciones.
	Fase 2: Análisis Exploratorio y Modelado
Mes 2	<ul style="list-style-type: none"> - Análisis exploratorio de datos (EDA). - Cálculo de densidades y brechas de oferta/demanda. - Desarrollo y entrenamiento del modelo de optimización de localización.
	Fase 3: Validación y Visualización
Mes 3	<ul style="list-style-type: none"> - Validación del modelo con datos de prueba. - Generación de mapas de calor y dashboards interactivos. - Elaboración del informe de resultados y recomendaciones.
	Fase 4: Documentación y Presentación
Mes 4	<ul style="list-style-type: none"> - Redacción final de la memoria técnica del proyecto. - Preparación de la presentación ejecutiva. - Socialización de resultados con stakeholders.

7. Relación con las 5V del Big Data

1. **Volumen (Volume):** Se procesarán datasets masivos que contienen registros de millones de personas y hogares, típico de un censo nacional, lo que requiere almacenamiento y procesamiento distribuido.
2. **Variedad (Variety):** Se integran datos estructurados (tablas del censo), semi-estructurados (JSON/XML de metadatos) y no estructurados

(direcciones en texto libre para geocodificar), además de datos geoespaciales (coordenadas, polígonos).

3. **Velocidad (Velocity):** Aunque el censo es estático, el procesamiento y análisis de estos grandes volúmenes de datos requiere una alta velocidad de cómputo para ser práctico, utilizando procesamiento en paralelo.
4. **Veracidad (Veracity):** La calidad y confiabilidad de los datos son cruciales. Se trabajará con fuentes oficiales (Institutos de Estadística) y se dedicará una fase completa a la limpieza y validación de los datos para asegurar la calidad de los resultados.
5. **Valor (Value):** El valor final es extremadamente alto: la capacidad de transformar datos brutos en conocimiento accionable para una óptima planificación educativa, con impactos sociales y económicos tangibles.

Arquitectura Técnica (Con Diagrama Explicativo)

El flujo de datos se estructuró en tres etapas:

- 1) Adquisición, donde se recopilaron datos censales, educativos y geoespaciales desde portales oficiales.
- 2) Transformación, donde se ejecutaron procesos para limpieza, depuración de valores atípicos, estandarización de direcciones y geocodificación.
- 3) Análisis, donde los datos integrados fueron utilizados para modelos espaciales y de optimización. El siguiente diagrama resume el proceso:

Insights Clave: 3 Hallazgos Principales

1. Alta concentración de población escolar por departamentos: El análisis espacial reveló áreas con crecimiento acelerado donde la infraestructura educativa no ha crecido a la misma velocidad.
2. Desbalance entre oferta y demanda en entre unidades educativas: Algunos departamentos presentan centros educativos subutilizados mientras otros muestran saturación extrema.
3. Departamentos óptimos detectados: Los modelos p-mediana y densidad indicaron coordenadas específicas que minimizan la distancia promedio de los estudiantes a un centro educativo cercano.