

Proyecto integrador — Electiva Big Data

Maestría en Estadística Aplicada · Universidad de Nariño · A-2026

1. De qué trata el proyecto

Ustedes y su grupo forman un equipo de analítica que apoya decisiones de política pública en Colombia. El trabajo del curso es construir un pipeline de big data reproducible: Access, Assess y Address sobre datos reales.

La pregunta concreta, los interesados y el subconjunto de datos los define cada grupo. Este documento fija el marco común y las tres opciones de enfoque para el proyecto.

2. Datos disponibles

Dataset obligatorio: microdatos GEIH del DANE (Gran Encuesta Integrada de Hogares). En la Lección 1 acceden a archivos mensuales; a lo largo del curso trabajan con particiones armonizadas incluyendo rangos de años mayores (p. ej. 2022–2025).

El dataset GEIH trae información de mercado laboral a nivel persona-mes: empleo, informalidad, ocupación, ingresos (según año), geografía (departamento, área), pesos de expansión y otras variables demográficas relevantes para el proyecto.

Las preguntas de decisión deben responderse con agregados del dataset GEIH (departamento, segmento, periodo). No se orienta el proyecto a identificar personas ni a usar datos que la encuesta no contiene.

3. Fuentes adicionales (opcional, recomendado)

Animamos a enriquecer el pipeline con otras fuentes abiertas cuando aporten a la pregunta de decisión. Ejemplos previstos en el curso: OpenStreetMap (contexto geográfico: servicios, vías, POI agregados por zona), Tablas administrativas abiertas del DANE u otras entidades públicas. El dataset GEIH sigue siendo la base común del curso. Las fuentes extra complementan, no la reemplazan.

4. Tres opciones de enfoque (elija uno por grupo)

Todos los grupos usan el dataset GEIH. Lo que cambia es el tipo de decisión que apoyan. Escoja una de las tres opciones de enfoque.

A. Asignación de recursos

Distribuir programas, servicios o inversión entre regiones o segmentos para mejorar el mercado laboral.

- Ejemplo: ¿Qué departamentos o segmentos de población deberían ser prioridad para un programa de empleo este año?
- Indicadores típicos: tasas de informalidad, desempleo o inactividad por departamento, con pesos de expansión.
- OpenStreetMap (opcional): agregar por departamento o municipio la densidad de servicios (salud, educación, transporte) para contrastar prioridades GEIH con contexto geográfico.
- Límite: el dataset GEIH informa prioridades, no registra cupos ni presupuestos de programas.

B. Riesgo / alerta temprana

Identificar zonas o condiciones donde el riesgo laboral aumenta para tomar medidas preventivas.

- Ejemplo: ¿Dónde aumentó más la informalidad o el desempleo entre un periodo definido?
- Indicadores típicos: cambios en el tiempo por departamento o segmento.
- OpenStreetMap (opcional): señalar zonas donde empeoran indicadores GEIH y faltan vías o POI de apoyo (servicios, conectividad) según mapas agregados.
- Límite: el dataset GEIH es encuesta mensual con rezago, no alerta en tiempo real. Formule la pregunta como tendencia o deterioro relativo, no predicción instantánea.

C. Monitoreo / seguimiento

Seguir un indicador en el tiempo y reportarlo a decisores.

- Ejemplo: ¿Cómo evolucionan las brechas regionales en participación laboral entre un periodo definido?
- Indicadores típicos: series anuales o mensuales comparables, con armonización de códigos entre años.
- OpenStreetMap (opcional): fijar una capa geográfica de referencia (p. ej. POI o red vial por zona) para interpretar tendencias GEIH en el mismo territorio.

5. Pregunta de referencia del curso

Independiente del enfoque, en los entregables finales cada grupo debe responder al menos una vez:

«¿Qué dice nuestro pipeline sobre patrones del mercado laboral para un segmento de población definido, y qué recomendaríamos o no recomendaríamos con base en ello?»

6. Próximos pasos

1. Lea este documento con su grupo y discútanlo durante las sesiones grupales.
2. Elija un enfoque y redacten `project_requirements.pdf` con la plantilla Word del curso. Este documento se discutirá en la segunda semana del curso.
3. Si planean usar un dataset adicional, documentenlo en `project_requirements.pdf`.
4. Alimente la definición y desarrollo del proyecto de acuerdo a lo explorado en las lecciones del curso.