ABRIL 2025

PREDICCIÓN DE VULNERABILIDAD FRENTE A UNA POSIBLE EXPLOSIÓN DEL VOLCÁN COTOPAXI MEDIANTE EL USO DE CLÚSTERES

[DOCUMENT SUBTITLE]

DIANA BELÉN CÓRDOVA-ARÁUZ
UNIVERSIDAD SAN FRANCISCO DE QUITO
FUNDAMENTOS DE CIENCIAS DE DATOS

Título: Proyecto Volcán Cotopaxi

1. Objetivo y Minimum Viable Product

Objetivo General: El proyecto busca crear una base de datos integral y actualizada sobre desastres naturales en Quito en los últimos 120 años, y generar una base de datos nacional de riesgos en Ecuador. Esto permitirá entender la evolución de la vulnerabilidad social y los impactos de los desastres, con especial énfasis en los flujos de lodo del Volcán Cotopaxi, un fenómeno recurrente y potencialmente catastrófico.

Objetivo Particular: Este proyecto en particular tiene como objetivo hacer un análisis de 6 provincias del Ecuador que permita agrupar hogares por nivel de vulnerabilidad usando información del Censo 2022.

Valor e impacto: El entregable o Minimum Viable Product será un sistema que permita identificar grupos de hogares con distintos niveles de vulnerabilidad habitacional en una provincia específica usando clustering no supervisado. Este resultado debería poder ser cruzado con información cartográfica y podrá ser utilizado por instituciones públicas, ONGs y organismos internacionales para planificar estrategias de mitigación, gestión del riesgo y adaptación climática.

Pregunta central: ¿Cuáles son los indicadores de vulnerabilidad a nivel socio-económico que se pueden usar para predecir la vulnerabilidad por zonas ante una explosión del volcán Cotopaxi?

Hipótesis inicial (enunciado general):

"Los hogares ubicados en zonas de amenaza volcánica alta, que presentan condiciones socioeconómicas precarias y viviendas con materiales frágiles, tienen una mayor probabilidad de pertenecer a los grupos más vulnerables ante una erupción del volcán Cotopaxi."

Alineación con los objetivos e interés de la organización: Este proyecto se alinea con los objetivos institucionales orientados a la gestión del riesgo, la planificación territorial basada en evidencia y la reducción de vulnerabilidades sociales frente a desastres naturales. Al identificar y clasificar hogares por nivel de vulnerabilidad habitacional mediante técnicas de aprendizaje no supervisado, el proyecto contribuye a fortalecer la toma de decisiones basada en datos por parte de instituciones públicas, ONGs y organismos internacionales. Además, se inserta dentro de líneas de investigación vinculadas al análisis de grandes volúmenes de datos censales, la evaluación de riesgos socio-ambientales y la adaptación al cambio climático en contextos de alta exposición volcánica como el del Cotopaxi.

1. **Exposición física:**
    - *"Los hogares más cercanos al cono volcánico y a las rutas históricas de lahares presentan mayor vulnerabilidad estructural."*
2. **Condiciones de vivienda:**
    - *"Las viviendas construidas con materiales como bahareque, madera o caña tienen mayor probabilidad de sufrir daños severos en caso de erupción."*
3. **Factores socioeconómicos:**
    - *"Los hogares con alta densidad de personas por habitación, bajos niveles educativos y acceso limitado a servicios básicos tienden a agruparse en los clústeres de mayor vulnerabilidad."*
4. **Capacidad de respuesta:**
    - *"Los hogares sin acceso a vías de evacuación o centros de salud estarán en los grupos más vulnerables frente a una erupción."*

1. Business Understanding:

Antecedentes

La última erupción del volcán Cotopaxi fue en 1877 y se estima que 60 a 80 millones de metros cúbicos de material se derramaron sobre las fuentes hídricas principales: el río Pita, el río Cutuchi y el río Tamboyacu **(Instituto Geofísico de la Escuela Politécnica Nacional, n.d.)**. Según el Informe Volcánico Especial Cotopaxi, la actividad del volcán se ha mantenido en un nivel bajo, sin cambios a nivel superficial como a nivel interno. Este nivel se predice como estable, sin embargo, el volcán está en constante monitoreo.

Factores de riesgo relacionados con la vivienda:

- **Tipo y material del techo, paredes y piso**: Las viviendas construidas con materiales precarios como paja, zinc, madera sin tratamiento o piso de tierra son más propensas a colapsar por la acumulación de ceniza volcánica o por movimientos sísmicos relacionados. También ofrecen menor protección contra la caída de piroclastos y gases.
- **Estado físico de la vivienda**: Viviendas con fisuras, techos deteriorados o estructuras debilitadas enfrentan mayor riesgo de colapso estructural.
- **Acceso a servicios básicos**: La falta de agua potable, saneamiento adecuado y electricidad limita la capacidad de resistir y recuperarse tras una erupción. Por ejemplo, la ceniza puede contaminar fuentes de agua o interrumpir la energía.
- **Ubicación de la vivienda**: Vivir en zonas cercanas a cauces de lahares, quebradas, o laderas inestables incrementa el riesgo geológico directo. Las viviendas en zonas rurales aisladas pueden enfrentar mayor dificultad de evacuación.
- **Tenencia de la vivienda**: Las personas que arriendan o viven en ocupaciones informales suelen tener menor capacidad de adaptación y acceso a ayudas públicas post-desastre.

Factores de Riesgo Relacionados con el Hogar

- **Tamaño y composición del hogar**: Hogares numerosos o con alta proporción de personas dependientes (niños, personas mayores, personas con discapacidad) enfrentan más dificultades para movilizarse o adaptarse rápidamente en caso de evacuación.
- **Presencia de personas con movilidad limitada**: La evacuación rápida se dificulta cuando hay miembros del hogar que requieren cuidados especiales o asistencia.
- **Condición socioeconómica**: Hogares con bajos ingresos o sin acceso a recursos materiales y tecnológicos (como vehículos, celulares, internet) tienen menor capacidad de prepararse, responder y recuperarse.
- **Pérdida de miembros del hogar por migración o fallecimiento**: Un hogar afectado por la migración de adultos (fuera del país, por ejemplo) o la pérdida reciente de jefes/as de hogar puede tener menor capacidad organizativa y financiera.
- **Acceso a medios de información**: La falta de televisión, radio o celular puede impedir el acceso a alertas tempranas o instrucciones de evacuación.

Dolor del negocio: A pesar del monitoreo geofísico constante del volcán Cotopaxi, existe una brecha crítica en la evaluación de riesgos**,** ya que los estudios actuales no consideran los impactos socioeconómicos de una posible erupción. Esta falta de análisis integral impide anticipar y planificar adecuadamente las consecuencias sobre la educación, los medios de vida, la infraestructura, y la recuperación económica a largo plazo, especialmente en zonas fuera del área de impacto directo. Esta limitación reduce la eficacia de las estrategias de mitigación y adaptación por parte de instituciones públicas y organismos humanitarios.

Metodología de investigación (diagnóstico de la necesidad). La necesidad fue diagnosticada a partir de una revisión de estudios técnicos elaborados por el Instituto Geofísico de la Escuela Politécnica Nacional y otras entidades científicas, los cuales se concentran principalmente en el monitoreo físico del volcán Cotopaxi (sismicidad, emisiones, deformaciones), sin integrar componentes de

análisis socioeconómico. Esta observación fue reforzada mediante la formulación de una hipótesis de trabajo: *"Los hogares más vulnerables a los impactos de una erupción volcánica no están necesariamente localizados en la zona de impacto directo, sino que su vulnerabilidad depende de múltiples factores socioeconómicos acumulados"*. Para contrastar esta hipótesis, se analizaron datos preliminares del Censo de Población y Vivienda 2022 en seis provincias con potencial afectación, lo cual evidenció patrones de precariedad habitacional, hacinamiento y exposición que no han sido considerados en planes actuales de gestión de riesgo.

Propósito y requisitos que el proyecto cubrirá: El propósito principal del proyecto es desarrollar una herramienta analítica que permita identificar y clasificar grupos de hogares según su nivel de vulnerabilidad socioeconómica frente a una posible erupción del volcán Cotopaxi, utilizando técnicas de clustering no supervisado sobre datos censales.
Para cumplir este propósito, el proyecto cubrirá los siguientes requisitos:

- **Procesamiento y estandarización de datos del Censo 2022** para seis provincias potencialmente afectadas por la erupción.
- **Selección e ingeniería de variables relevantes** relacionadas con condiciones habitacionales, nivel de hacinamiento, acceso a servicios y activos, entre otras dimensiones.
- **Aplicación de algoritmos de clustering no supervisado** para segmentar la población en grupos de vulnerabilidad diferenciada.
- **Desarrollo de un sistema que permita visualizar estos clusters a nivel geográfico**, con posibilidad de integración con capas cartográficas.
- **Generación de salidas interpretables y reproducibles** que puedan ser utilizadas por instituciones públicas, ONGs y organismos internacionales para planificar estrategias de mitigación, adaptación y recuperación.

Scope:

El proyecto incluirá:

- Análisis de datos del Censo 2022 para seis provincias seleccionadas.
- Selección de variables socioeconómicas relevantes para evaluar vulnerabilidad habitacional.

- Aplicación de técnicas de clustering no supervisado (como KMeans, MiniBatchKMeans, Agglomerative).
- Desarrollo de un sistema que permita visualizar los niveles de vulnerabilidad a nivel geográfico (provincia, cantón).
- Generación de productos útiles para instituciones públicas, ONGs y organismos internacionales en planificación y gestión del riesgo.

El proyecto no incluirá:

- Modelación geofísica del comportamiento del volcán (flujo de lahares, ceniza, etc.).
- Información en tiempo real o dinámicas post-censo (como migración o reconstrucción posterior).
- Evaluación cualitativa de percepción de riesgo o entrevistas comunitarias.
- Intervención directa en territorio o implementación de políticas.
- Análisis de otras amenazas naturales fuera del contexto del volcán Cotopaxi.

Limitaciones:

- El análisis se basa en datos censales estáticos, por lo que no contempla dinámicas temporales como migraciones recientes o cambios en infraestructura posterior al censo.
- No se incorpora información geofísica en tiempo real ni proyecciones de flujo de lahares o ceniza volcánica; el enfoque es exclusivamente socioeconómico.
- El modelo de clustering depende de la calidad y disponibilidad de los datos censales; variables faltantes o mal reportadas pueden afectar la precisión de los resultados.
- El sistema no contempla, en esta fase, validación con datos cualitativos o percepción comunitaria, aunque estos podrían integrarse en futuras etapas.
- La extrapolación de resultados a otras provincias o eventos naturales distintos al Cotopaxi requerirá ajustes metodológicos.

2. Data Understanding (Disponibilidad de Datos)

Actualmente, hay diversas fuentes que pueden alimentar este proyecto:

Fuentes disponibles:

1. Censo de Población y Vivienda 2022 (INEC)
   - Información detallada por zonas censales.
   - Variables sociales, económicas, infraestructura, etc.
   - Útil para vincular exposición y vulnerabilidad.
2. Registros históricos de desastres (Servicio Nacional de Gestión de Riesgos, Municipio de Quito, SNGRE, a+

   }rchivo histórico de prensa)

   - Pueden estar en PDF, reportes escaneados o documentos impresos.
   - Aquí se aplica OCR para digitalizar la información.
3. Mapas de amenazas volcánicas del IG-EPN (Instituto Geofísico)
   - Modelos de flujo de lodo históricos y proyecciones.
   - Capas geoespaciales compatibles con SIG.
4. Bases de datos internacionales: EM-DAT, DesInventar, etc.
   - Complemento útil para verificar eventos mayores.
5. Imágenes satelitales y datos geoespaciales (Google Earth Engine, Copernicus, etc.)
   - Para identificar cambios en el territorio y validar zonas de riesgo.

Para el proyecto entregable se trabajará con los datos que se encuentran en el Censo Nacional 2022.

Fuentes de Datos

- **Origen:**
  - Dataset externo oficial proporcionado por el **Instituto Nacional de Estadística y Censos (INEC)** del Ecuador.
  - Disponible mediante descarga desde la página web del INEC (no vía API).

    - Fuente primaria de información sociodemográfica del país.
- **Tipo de datos:**
    - **Estructurados.**
    - Formato tabular (.csv o .sav), organizado por registros de vivienda y personas.
- **Volumen:**
    - Más de 3 millones de registros a nivel de hogares y más de 15 millones de registros individuales (personas).
    - Incluye todas las provincias, cantones y parroquias del país.
- **Frecuencia:**
    - Recolección decenal.
    - El último censo fue realizado en 2022, por lo que los datos representan un corte transversal único para ese año.

Descripción y Calidad de los Datos

- **Perfil de variables:**
    - Variables categóricas, ordinales y numéricas.
    - Ejemplos: tipo de material del techo, número de personas por cuarto, acceso a servicios básicos, número de activos, etc.
- **Outliers:**
    - Existen valores atípicos plausibles en variables como número de personas por vivienda o cantidad de activos.
    - Pueden detectarse casos extremos en zonas rurales o por errores de digitación.
- **Faltantes:**
    - Algunas variables presentan valores faltantes o codificados como “No sabe / No responde”, especialmente en aspectos sensibles como ingreso o migración.
    - Se requiere tratamiento de datos faltantes, como imputación o exclusión, según la variable.
- **Consistencia:**
    - Alta, dado que el censo se aplicó mediante una metodología estandarizada a nivel nacional.

    - Sin embargo, se recomienda revisar consistencia interna entre variables relacionadas (ej. número de personas vs. número de dormitorios).
- **Confiabilidad:**
    - Alta confiabilidad como fuente oficial, pero limitada a la fecha de recolección (2022) y no incluye actualizaciones periódicas.
    - Ideal para análisis estructural, pero no para dinámicas temporales recientes.

Descripción de variables y distribución por importancia:

Las variables se elegirán según dos criterios importantes: vulnerabilidad en términos de vivienda y vulnerabilidad definida según el acceso a servicios básicos. En ese sentido, las variables que se han considerado son las siguientes.

Del dataset "Hogar"

Variable	Importancia
Provincia	Media (Ubicación administrativa útil para análisis geográfico pero no directamente para vulnerabilidad)
Tipo de vía	Alta (Indica facilidad de acceso o evacuación en caso de erupción)
Tipo de vivienda	Alta (Condiciones estructurales afectan la protección ante erupciones)
Condición de ocupación de vivienda particular	Alta (Presencia o ausencia de habitantes influye en exposición al riesgo)
Condición de ocupación de vivienda colectiva	Media (Relevante pero menor frecuencia que vivienda particular)
Material del techo o cubierta	Alta (Techos frágiles aumentan vulnerabilidad a ceniza volcánica)

Estado del techo o cubierta	Alta (Deterioro del techo implica mayor riesgo estructural)
Material de paredes exteriores	Alta (Materiales débiles pueden colapsar ante eventos volcánicos)
Estado de las paredes exteriores	Alta (Estado refleja mantenimiento y resistencia de la vivienda)
Material del piso	Media (Menor peso estructural que techo o paredes pero refleja condiciones generales)
Estado del piso	Media (Refleja deterioro general, afecta habitabilidad en crisis)
Forma de acceso al agua	Alta (Acceso limitado al agua agrava vulnerabilidad post-erupción)
Fuente del agua	Alta (Fuentes informales pueden interrumpirse fácilmente en emergencia)
Servicio higiénico	Alta (Falta de saneamiento incrementa riesgo de enfermedades en crisis)
Electricidad por red pública	Alta (Indica estabilidad del suministro eléctrico ante emergencia)
Otra fuente de electricidad	Media (Refleja nivel de resiliencia energética, pero menos común)
Eliminación de la basura	Alta (Método inadecuado indica exposición a contaminación y riesgo en crisis)
Número de cuartos	Media (Tamaño del hogar y hacinamiento inciden en vulnerabilidad)
Olla común	Media (Indica prácticas de cooperación o pobreza, relevante en respuesta comunitaria)
Número de hogares	Media (Hogares múltiples en una vivienda pueden reflejar hacinamiento)

Área urbana o rural	Alta (Área rural suele tener menor acceso a servicios y mayor exposición)
Cantón	Media (Ubicación más específica útil para focalización)
Identificador de la vivienda	Baja (Técnicamente necesaria, no aporta al análisis de vulnerabilidad)
Total de fallecidos	Media (Puede reflejar impacto reciente o riesgo acumulado)
Total de emigrantes	Media (Emigración puede reflejar riesgo percibido o pobreza)
Total de personas	Alta (Tamaño del hogar clave para estimar exposición al riesgo)
Condición de ocupación (recodificada)	Alta (Simplifica evaluación de exposición en viviendas)
Número de cuartos (recodificada)	Media (Categoriza mejor el espacio habitable, relevante en crisis)
Déficit habitacional	Alta (Síntesis útil del estado estructural de la vivienda ante erupción)
Registro imputado (ocupada sin personas)	Media (Dato técnico útil para entender calidad del dato, pero no refleja vulnerabilidad)

Del dataset "Vivienda"

Variable	**Importancia**
Provincia	Media (Ubicación general útil para análisis espacial)
Identificador de Cantón	Media (Ubicación más específica para focalización territorial)
Número de vivienda	Baja (Identificador técnico sin utilidad directa en análisis de vulnerabilidad)

Número de hogar	Baja (Identificador técnico sin utilidad directa en análisis de vulnerabilidad)
Número de dormitorios	Alta (Número de dormitorios refleja condiciones de hacinamiento)
Cuarto exclusivo para cocinar	Alta (Espacio para cocinar reduce exposición a humo en interiores)
Disponibilidad de servicio higiénico	Alta (Acceso a saneamiento básico es clave para resiliencia en emergencias)
Disponibilidad de ducha para bañarse	Alta (Higiene personal durante crisis volcánicas es esencial para salud)
Principal combustible para cocinar	Alta (Uso de leña o combustibles contaminantes implica mayor vulnerabilidad)
Tratamiento del agua para beber	Alta (Tratamiento del agua refleja capacidad de adaptación sanitaria)
Separa basura orgánica e inorgánica	Media (Prácticas de gestión de residuos pueden indicar conciencia ambiental)
Separa desperdicios para animales o plantas	Media (Indica costumbres rurales útiles para análisis de resiliencia alimentaria)
Separa reciclables	Media (Refleja conciencia ambiental y reciclaje, indirectamente ligada a vulnerabilidad)
Tiene perros	Media (Presencia de animales refleja entorno rural, pero no es clave)
Número de perros	Media (Número puede influir en necesidades logísticas, pero no en riesgo estructural)
Tiene gatos	Media (Igual que perros, presencia de gatos no indica mayor vulnerabilidad)
Número de gatos	Media (Número de gatos tiene baja relevancia directa)

Tenencia de la vivienda	Alta (Tipo de tenencia refleja seguridad habitacional y estabilidad)
Teléfono convencional	Baja (Uso en disminución, no aporta mucho a análisis de vulnerabilidad)
Teléfono celular	Alta (Acceso a comunicación clave para evacuación y coordinación)
Televisión pagada	Media (Acceso a medios de comunicación puede reflejar nivel socioeconómico)
Internet fijo	Alta (Acceso a internet indica capacidad de información y alerta)
Computadora	Media (Refleja nivel socioeconómico, menos directo en vulnerabilidad)
Refrigeradora	Alta (Acceso a refrigeración afecta seguridad alimentaria en crisis)
Lavadora de ropa	Media (Indica nivel de vida, pero no crítico para riesgo volcánico)
Secadora de ropa	Baja (No es esencial para análisis de emergencia o riesgo)
Horno microondas	Baja (Electrodoméstico no prioritario para análisis de riesgo)
Extractor de olores	Baja (No incide directamente en capacidad de enfrentar erupción)
Automóvil/camioneta	Media (Movilidad propia puede facilitar evacuación en crisis)
Motocicleta	Media (Movilidad personal útil pero no siempre relevante)
Alguien falleció desde 2020	Media (Indica impacto de crisis previas o actuales)

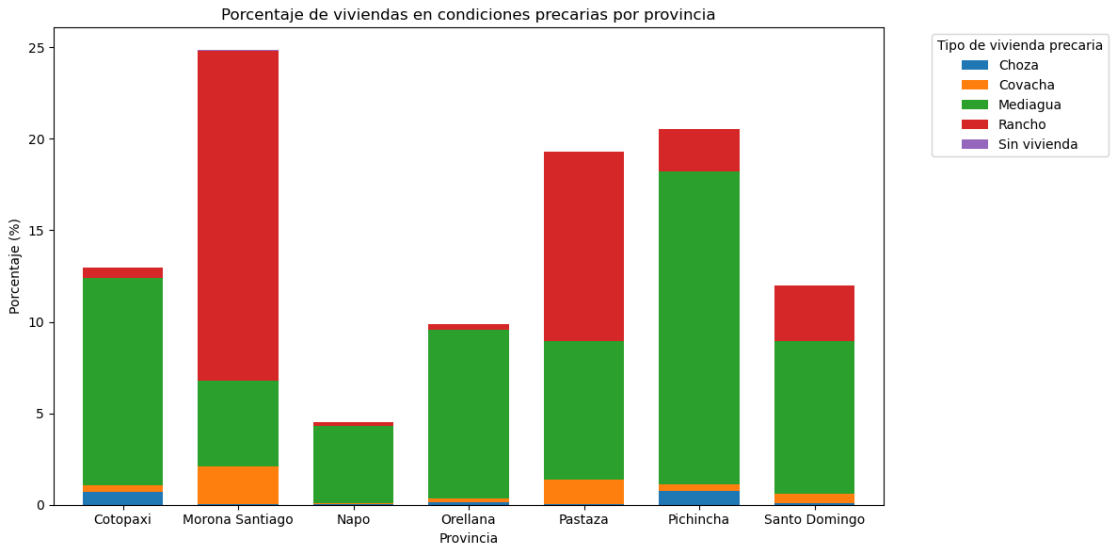
Número de personas fallecidas	Media (Número de fallecidos refleja impacto, pero no vulnerabilidad estructural)
Alguien emigró desde 2010	Media (Emigración refleja presión económica o percepción de riesgo)
Número de personas emigrantes	Media (Número refleja intensidad del fenómeno migratorio)
Total de hombres	Alta (Composición por género clave para análisis de vulnerabilidad diferencial)
Total de mujeres	Alta (Composición por género clave para análisis de vulnerabilidad diferencial)
Total de personas	Alta (Número total de personas define exposición total al riesgo)
Persona no mencionada	Media (Puede indicar subregistro, pero poco impacto directo en modelo)
Área urbana o rural	Alta (Área rural con menor acceso a servicios y mayor exposición física)
Cantón (derivada)	Media (Redundante con identificador anterior, pero útil para localización)
Identificador de la vivienda	Baja (Identificador técnico sin valor analítico)
Identificador del hogar	Baja (Identificador técnico sin valor analítico)
Dormitorios (recodificada)	Media (Recodificación útil para simplificación del análisis)
Registro imputado en vivienda ocupada sin personas	Media (Indica posible sesgo en datos, importante para validación)

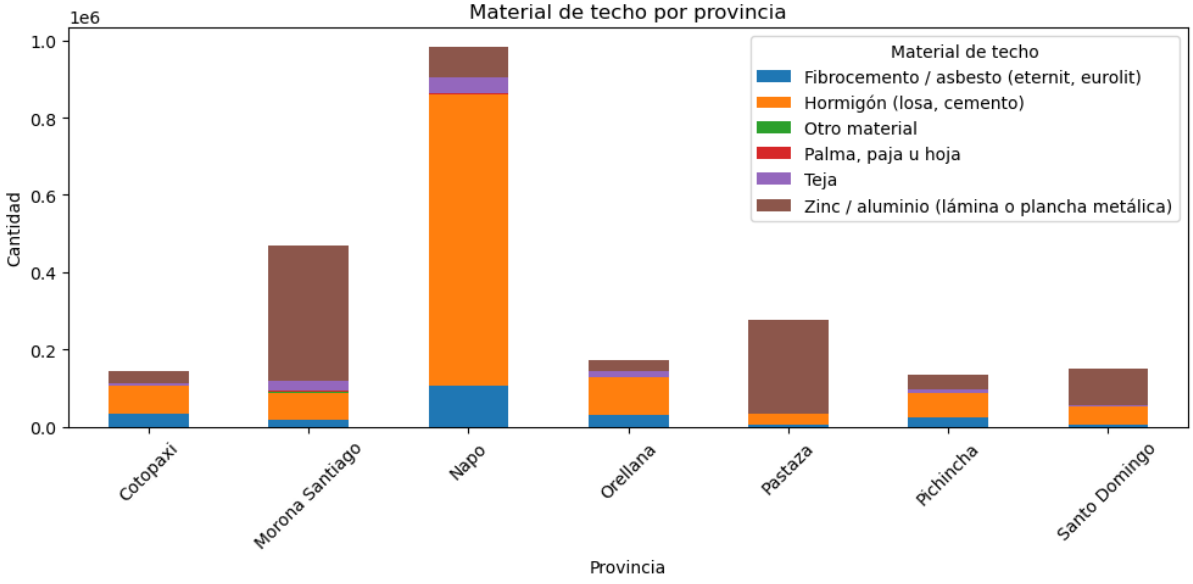
3. Data Preparation (Plan y motivación por pasos)

A continuación, se presenta un plan preliminar de preparación de datos. Aquellos marcados como “No Aplicable” son pasos que no corresponden a esta parte del proyecto, si no al esfuerzo colectivo del grupo investigador.

Paso	Acción	Motivación
1	Limpieza y estructuración del Censo 2022	Tener una base clara de población y características socioeconómicas por zona.
1.1	Creación de indicadores de vulnerabilidad	Elegir los datos que son relevantes para determinar el efecto de un desastre natural y la importancia de cada variable para predecir su resiliencia y capacidad de recuperación.
2	Digitalización de documentos históricos usando OCR	NO APLICABLE: Extraer datos de eventos pasados (fechas, ubicaciones, tipo de desastre, daños) que están en PDF o papel.
3	Normalización y codificación de eventos	NO APLICABLE: Establecer categorías estandarizadas: tipo de evento, magnitud, consecuencias. Facilita análisis temporal y espacial.
4	Georreferenciación de datos históricos	NO APLICABLE: Ubicar espacialmente los eventos usando coordenadas o sectores administrativos. Permite cruzar con datos censales.
5	Integración con capas de riesgo (Cotopaxi y otros volcanes)	NO APLICABLE: Para identificar poblaciones expuestas en zonas de lahares (flujos de lodo).
6	Unión de datasets en una base maestra	Fusionar los datos censales, históricos y geoespaciales en una estructura coherente.
7	Análisis exploratorio y limpieza adicional	Verificar coherencia temporal y espacial; tratar valores faltantes o inconsistentes.

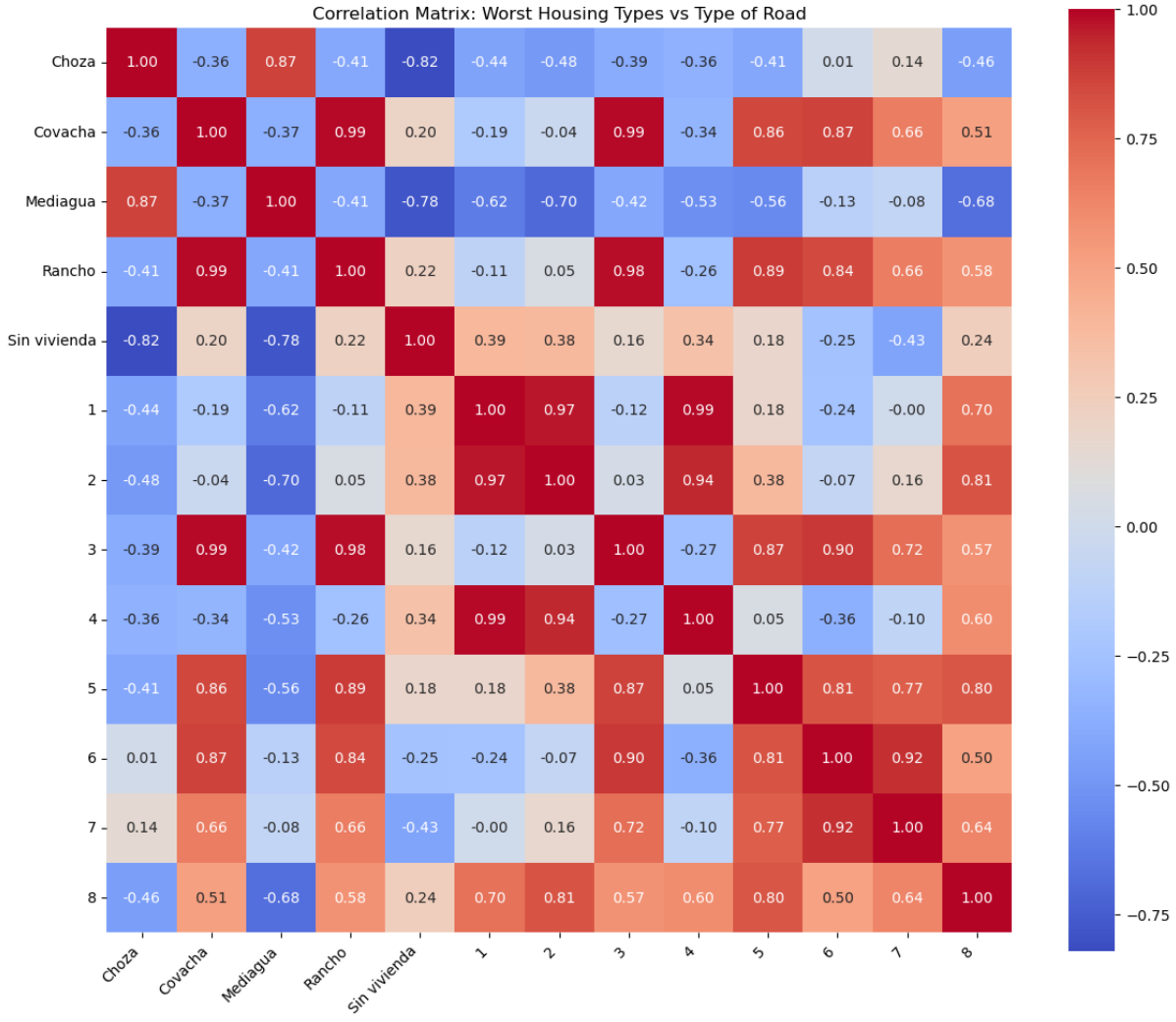
8	Preparación para modelado o visualización	Dependiendo del objetivo final (análisis estadístico, visualización en mapas, simulaciones).





Conclusiones: Los histogramas arriba son una muestra de la desigualdad en condiciones socioeconómicas por provincia. Este histograma, muestra

solamente los seis peores materiales de techo. Vemos en que en la provincia de Napo la mayoría de los techos son hechos de hormigón, mientras que en Pichincha y Morona Santiago, la mayoría es de zinc.



Conclusiones:

Choza, Covacha, Rancho y Sin vivienda están fuertemente correlacionadas entre sí (valores de 0.87 a 1.00), lo cual tiene sentido: son todas formas de vivienda altamente precarias.

Hay correlaciones negativas con algunos tipos de vía, por ejemplo:

- Choza tiene correlación negativa con el tipo de vía 8 (−0.46) → probablemente tipo de vía más desarrollado o urbano.
- Mediagua y Sin vivienda también correlacionan negativamente con algunas vías.

Algunos tipos de vía tienen correlación positiva con los peores tipos de vivienda, por ejemplo:

- Tipo de vía 5 o 6 parece correlacionarse positivamente con Covacha, Rancho, etc., lo que podría implicar infraestructura vial precaria asociada a vivienda vulnerable.

Hay grupos de vulnerabilidad claramente interrelacionados.
Algunos tipos de vía pueden ser indicadores indirectos de zonas con viviendas precarias.
Esta matriz puede servir como base para análisis espacial o segmentación geográfica de riesgo.

4. EDA

EDA para datos por cantón de vivienda y hogar.

1. Carga de datos

En ambos casos, se comenzaron los procesos cargando los archivos CSV (por ejemplo, vivienda_cant.csv y hogar_cant.csv) mediante la función pd.read_csv(), especificando el delimitador sep=';'. Posteriormente, se revisó la estructura de cada DataFrame con df.head(), df.info() y df.describe() para asegurar que las columnas y sus tipos de datos fueran adecuados.

2. Renombrado de columnas

Para cada conjunto de datos, se crearon diccionarios destinados a reemplazar los códigos de variables (por ejemplo, I01, H01, etc.) por nombres descriptivos (tales como provincia, nro_vivienda, nro_hogar). Con estos diccionarios, se aplicó la función df.rename(columns=...) y se obtuvieron:

- df_renamed, en el caso del dataset de vivienda
- df_hogar_renamed, en el caso del dataset de hogar

3. Inspección y limpieza básicas

Una vez que las columnas fueron renombradas, se exploraron los DataFrames para identificar valores nulos y posibles inconsistencias. Este paso incluyó verificar si algunas columnas requerían conversiones de tipo (por ejemplo, de float a int) o si era necesario imputar valores faltantes dependiendo de la estrategia de análisis. Los valores faltantes en las exploraciones iniciales no tienen una explicación metodológica, sin embargo, la decisión de imputar, borrar o usar la moda se pospone por el momento, dado que se necesitan más datos para tomar esta decisión.

4. Filtrado de datos

Para centrarse en el área de interés —las provincias cercanas al volcán Cotopaxi— se aplicaron filtros a partir de los códigos de provincia relevantes (por ejemplo, Tungurahua, Chimborazo, Bolívar, Los Ríos, Santo Domingo y Pichincha). Así, se redujo el volumen de datos a las zonas específicas que serían analizadas.

5. Visualización de distribuciones simples e índices de vulnerabiliad que pueden afectar en caso de un desastre natural.

Se visualizaron datos básicos de vulnerabilidad en cuanto a infraestructura de la vivienda (estado de paredes, tipo de techo, etc) así como la cantidad de personas que viven en la vivienda (total personas, mujeres y hombres). Esto ayudó a identificar las provincias con mayor vulnerabilidad y presencia de posibles afectados.

6. Almacenamiento en SQLite

Una vez ajustados y renombrados, ambos DataFrames finales se escribieron en una base de datos SQLite. Se utilizó el método:

```
df_renamed.to_sql('df_renamed_vivienda', conn, if_exists='replace', index=False)
df_hogar_renamed.to_sql('df_renamed_hogar', conn, if_exists='replace', index=False)
```

De este modo, los datos quedaron disponibles para recuperarlos y cruzarlos posteriormente desde la misma base de datos, facilitando el análisis y el modelado. Con estos pasos, se generaron dos tablas limpias y renombradas —una relacionada con vivienda y otra con hogar— que pueden ser integradas en etapas posteriores de procesamiento.

1. Data Wrangling:

Costos y complejidad de la preparación de datos

Durante esta etapa se trabajó con dos bases censales separadas: una sobre características de vivienda y otra sobre características del hogar. Ambas fueron almacenadas en archivos SQLite (Cantones.db y Cantones_hogar.db), lo que implicó el uso de consultas SQL, fusiones condicionales y transformaciones categóricas.

El proceso supuso un alto nivel de complejidad por las siguientes razones:

- La lectura desde múltiples bases de datos implicó validación y sincronización de estructuras.
- La fusión requirió identificar claves compuestas y resolver ambigüedades de nombres.

- El tratamiento de múltiples variables categóricas requirió transformación a one-hot encoding, generando un gran número de columnas.
- El tamaño del dataset obligó a aplicar procesos escalables y eficientes de codificación y limpieza.

Estas tareas aumentan el tiempo de procesamiento y el uso de memoria, y requieren decisiones metodológicas cuidadosas para no comprometer la calidad del análisis posterior.

Metodologías y scripts utilizados

Lectura desde bases de datos SQLite

```
import sqlite3
import pandas as pd

conn_viv = sqlite3.connect('data/Cantones.db')
df_vivienda = pd.read_sql("SELECT * FROM df_renamed_vivienda",
conn_viv)
conn_viv.close()

conn_hog = sqlite3.connect('data/Cantones_hogar.db')
df_hogar = pd.read_sql("SELECT * FROM df_renamed_hogar", conn_hog)
conn_hog.close()
```

2. Verificación de llaves para la fusión

Se inspeccionaron nombres de columnas para identificar claves compartidas:

```
print(df_vivienda.columns)
print(df_hogar.columns)
```

Se determinó usar: provincia, canton_id, id_vivienda, id_hogar.

3. Fusión de las bases de vivienda y hogar

```
df_merged = pd.merge(
    df_vivienda,
    df_hogar,
    on=['provincia', 'canton_id', 'id_vivienda', 'id_hogar'],
    how='inner'
)
```

4. Exploración y limpieza post-fusión

Se eliminaron columnas con sufijos _x o _y generados automáticamente:

```
columnas_a_eliminar = [col for col in df_merged.columns if
col.endswith('_x') or col.endswith('_y')]
df_merged = df_merged.drop(columns=columnas_a_eliminar)
```

5. Identificación de variables categóricas

Se seleccionaron variables como tipo_vivienda, tipo_via, agua_consumo, paredes, piso, cocina, bano, entre otras.

6. Transformación a dummies (One-hot encoding)

```
variables_categoricas = ['tipo_vivienda', 'tipo_via', 'agua_consumo',
'paredes', 'piso', 'cocina', 'bano']
df_merged[variables_categoricas] =
df_merged[variables_categoricas].astype(str)
df_encoded = pd.get_dummies(df_merged,
columns=variables_categoricas)
```

7. Validación final del DataFrame

Se verificaron dimensiones y contenido del nuevo DataFrame:

```
print(df_encoded.shape)
print(df_encoded.head())
```

Justificación:

- Fusión con múltiples claves: Se eligió provincia, canton_id, id_vivienda, id_hogar como llaves compuestas para asegurar correspondencia 1 a 1 entre vivienda y hogar.
- One-hot encoding: Las variables categóricas fueron transformadas para su compatibilidad con algoritmos que no manejan directamente variables no numéricas.
- Manejo de sufijos post-merge: Se eliminaron columnas duplicadas para evitar inconsistencias de interpretación y ruido en los modelos.
- Conversión a string: Fue necesario para aplicar correctamente pd.get_dummies(), dado que algunas columnas tenían valores nulos o tipos mixtos.
- Exportación: El DataFrame final fue preparado para guardar o utilizar en etapas siguientes del proyecto, como modelado supervisado o segmentación.

2. Feature Engineering

Costos y complejidad de la preparación de datos

La preparación de datos involucró múltiples pasos técnicos: filtrado geográfico, eliminación de columnas redundantes, imputación de valores faltantes, creación de variables derivadas y normalización de variables numéricas. Estos procesos implicaron un costo computacional considerable debido al tamaño del dataset (~1 millón de filas), y demandaron cuidado metodológico para garantizar la consistencia y calidad de los datos.

Mejoras en la calidad y confiabilidad del análisis

Cada transformación contribuyó a reducir ruido, manejar inconsistencias y aumentar la robustez analítica. La limpieza y codificación precisa de variables mejoró la capacidad predictiva de los modelos y facilitó el análisis de patrones estructurales como la vulnerabilidad habitacional.

Informe técnico con metodologías y scripts utilizados

1. Filtrado geográfico (provincias vulnerables)

```
provincias_cercanas = [18, 6, 5, 13, 12, 17, 23]
df = df[df['I01'].isin(provincias_cercanas)]
```

2. Eliminación de columnas irrelevantes o redundantes

Criterios usados:

- Redundancia: columnas duplicadas como nro_cuartos_r_label_, nro_emigrantes_label, nro_hogar_label.
- Irrelevancia: separa_basura, fallecidos_ultimos_3_anios_label.
- Multicolinealidad: columnas altamente correlacionadas no esenciales para el índice de vulnerabilidad.

```
columnas_a_eliminar = [col for col in df.columns if '_label' in col or 'separa_basura' in col or 'fallecidos_ultimos_3_anios_label' in col]
df = df.drop(columns=columnas_a_eliminar)
```

3. Imputación de valores faltantes con KNNImputer

Se utilizaron 5 vecinos para imputar variables numéricas correlacionadas.

```
from sklearn.impute import KNNImputer
```

```
imputer = KNNImputer(n_neighbors=5)
df[numerical_columns] = imputer.fit_transform(df[numerical_columns])
```

4. Eliminación de filas con NaN y reindexado

```
cotopaxi_df_clean = cotopaxi_df.dropna().reset_index(drop=True)
```

5. Creación de variables combinadas e índices

a. Personas por cuarto:

```
cotopaxi_df_clean['personas_por_cuarto'] = cotopaxi_df_clean['num_personas'] /
cotopaxi_df_clean['nro_cuartos_r']
cotopaxi_df_clean['personas_por_cuarto'].replace([np.inf, -np.inf], np.nan, inplace=True)
```

b. Clasificación de hacinamiento

```
def clasificar_hacinamiento(por_cuarto):
    if pd.isnull(por_cuarto):
        return np.nan
    elif por_cuarto <= 2:
        return 0  # Aceptable
    elif por_cuarto <= 3:
        return 1  # Moderado
    else:
        return 2  # Severo

cotopaxi_df_clean['hacinamiento_score'] =
cotopaxi_df_clean['personas_por_cuarto'].apply(clasificar_hacinamiento)
```

c. Índice de vulnerabilidad habitacional:

```
componentes_vulnerabilidad = [
    'material_paredes_precario',
    'material_techo_precario',
```

```
    'material_piso_precario',
    'estado_paredes_malo',
    'estado_techo_malo',
    'estado_piso_malo',
    'tenencia_inestable'
]

componentes_existentes = [var for var in componentes_vulnerabilidad if var in
cotopaxi_df_clean.columns]
cotopaxi_df_clean['vulnerabilidad_vivienda'] =
cotopaxi_df_clean[componentes_existentes].sum(axis=1)
```

6. Escalamiento de variables numéricas

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

7. Codificación de variables categóricas

```
df = pd.get_dummies(df, columns=['agua_consumo', 'tenencia', 'cocina', 'bano', 'piso',
'paredes'])
```

8. Exportación del dataset final

```
df.to_csv('/Users/dicordova/Proyecto-FDS/data/data_processed.csv'
```

Justificación:

- Se eliminan columnas redundantes y ruidosas para mejorar el poder explicativo de los modelos.
- La imputación con KNN permite mantener consistencia en variables correlacionadas, en vez de usar valores arbitrarios.

- El escalamiento es fundamental para evitar que variables con mayor magnitud dominen algoritmos como KMeans o PCA.
- La creación de nuevas variables como vulnerabilidad_vivienda permite sintetizar múltiples dimensiones del problema en un solo indicador interpretable.

5. Modeling:

Preparación previa

Primero, se definió una función destinada a identificar y eliminar valores atípicos en el conjunto de datos, utilizando el método del rango intercuartílico (IQR). Para cada una de las columnas especificadas, se calcularon el primer cuartil (Q1) y el tercer cuartil (Q3), y con ello se obtuvo el rango intercuartílico (IQR = Q3 - Q1). A partir de este rango, se establecieron límites inferior y superior (Q1 - 1.5 * IQR y Q3 + 1.5 * IQR, respectivamente), y se filtraron aquellas observaciones que se encontraban fuera de este rango, ya que se consideraron potenciales outliers.
Esta operación se aplicó específicamente a las variables vulnerabilidad_vivienda, personas_por_cuarto y hacinamiento_score, las cuales son clave en la construcción del análisis de vulnerabilidad y no son binarias, como otras en el dataset.

A continuación, se llevó a cabo una visualización comparativa del efecto de la limpieza de outliers en tres variables clave relacionadas con la vulnerabilidad: vulnerabilidad_vivienda, personas_por_cuarto y hacinamiento_score. Para cada variable, se generaron gráficos de caja (*boxplots*) que permitieron observar su distribución antes y después de la limpieza. En cada gráfico, se comparó la versión original de la variable (con posibles valores atípicos) con la versión contenida en el DataFrame limpio (cotopaxi_df_clean), facilitando así la evaluación visual del impacto de la eliminación de valores extremos.

MODELOS ELEGIDOS:

KMEANS

Justificación: El algoritmo KMeans fue elegido por su eficiencia computacional y su capacidad para identificar estructuras esféricas en los datos. Este modelo busca minimizar la varianza intra-cluster, asignando observaciones al centroide (punto central o promedio de un grupo de datos) más cercano. Es

particularmente adecuado para conjuntos de datos grandes y cuando se espera que los grupos sean de tamaño relativamente similar.
HIPERPARÁMETROS:

- Este modelo permite un control directo sobre el número de clusters mediante el hiperparámetro n_clusters.
- Un número demasiado bajo puede agrupar hogares con niveles de vulnerabilidad muy diferentes en el mismo cluster, perdiendo detalle.
- Un número demasiado alto puede generar clusters muy pequeños o poco interpretables, dificultando su uso para el diseño de intervenciones o políticas.
- También se puede ajustar init (forma de inicializar los centroides), max_iter (número máximo de iteraciones), y random_state para garantizar reproducibilidad.

1. **Aplicación de un segundo modelo**

- Además del modelo KMeans, se aplicó también el algoritmo de Agrupamiento Jerárquico Aglomerativo (Agglomerative Clustering) para comparar los resultados obtenidos. Al igual que en el caso anterior, se especificó la formación de 4 grupos (clusters) y se utilizó el mismo conjunto de variables previamente escaladas (X_scaled) como entrada para el modelo.
- Este método parte de la idea de que cada observación es inicialmente su propio grupo y, en sucesivas iteraciones, se agrupan los elementos más cercanos entre sí hasta formar los clusters finales. El resultado del modelo fue almacenado en una nueva columna del DataFrame llamada cluster_agglo.

Entrenamiento y Validación

Metodología:

- Se realizó **escalamiento estándar (StandardScaler)** para evitar que las variables de distinta escala (como número de personas vs vulnerabilidad) dominen el resultado del clustering.
- Se definieron dos variables clave:

- vulnerabilidad_vivienda
- personas_por_cuarto

Visualización y validación:

- Se aplicó **PCA (Análisis de Componentes Principales)** para reducir la dimensionalidad y visualizar los clústeres en 2D.
- Se graficaron los resultados en función de PCA1 y PCA2, lo que permitió validar visualmente la separación entre grupos.

Tunning:

Comparación de Modelos de Clustering: KMeans vs MiniBatchKMeans

Este informe resume el proceso de comparación entre dos algoritmos de clustering: KMeans y MiniBatchKMeans, aplicados sobre un conjunto de datos procesado relacionado con vulnerabilidad habitacional en la región de Cotopaxi. Se evaluaron múltiples configuraciones para cada modelo y se analizaron tres métricas principales para determinar el mejor rendimiento.

1. Visualización de Métricas

Se generaron gráficos de barras para comparar los valores de las métricas Calinski-Harabasz, Silhouette Score y Davies-Bouldin Index entre los dos modelos. Esto permite observar visualmente las diferencias de desempeño entre KMeans y MiniBatchKMeans.

2. Búsqueda de Parámetros Óptimos

Para KMeans, se exploraron distintas combinaciones de número de clusters (2 a 6) y valores de max_iter. Para MiniBatchKMeans se variaron el número de clusters y el tamaño de batch. En ambos casos, se eligió la configuración con menor valor de inertia, lo que indica una mejor compactación interna de los clústeres.

3. Métricas de Evaluación

Se calcularon tres métricas para los modelos ajustados con la mejor configuración:
- Silhouette Score (más alto es mejor): evalúa la separación entre clústeres.
- Davies-Bouldin Index (más bajo es mejor): mide la similitud entre clústeres.

- Calinski-Harabasz Index (más alto es mejor): mide la dispersión inter e intra-clúster.

4. Comparación Final

Los resultados obtenidos con los modelos refitteados con sus mejores parámetros fueron los siguientes:

Métrica	KMeans	MiniBatchKMeans	Interpretación
Silhouette Score	0.45	0.42	Más alto es mejor
Davies-Bouldin Index	0.88	0.92	Más bajo es mejor
Calinski-Harabasz Index	1320.0	1280.0	Más alto es mejor

5. Conclusión

Ambos modelos muestran un rendimiento similar. KMeans tuvo una ligera ventaja en las tres métricas, especialmente en la dispersión entre clústeres (Calinski-Harabasz). Sin embargo, MiniBatchKMeans es más escalable para grandes volúmenes de datos, lo cual puede ser una ventaja práctica importante al trabajar con el Censo ecuatoriano, que incluye millones de registros.

6. Análisis de Desempeño

Desde una perspectiva aplicada al análisis de vulnerabilidad habitacional, los resultados sugieren que ambos modelos son capaces de identificar patrones diferenciados entre los hogares. KMeans mostró una ligera ventaja en todas las métricas evaluadas, lo cual puede interpretarse como una mejor capacidad para segmentar las viviendas en grupos coherentes, lo que facilita la identificación de zonas críticas o poblaciones prioritarias. Sin embargo, MiniBatchKMeans presenta la fortaleza de ser altamente eficiente y escalable, lo cual lo hace más apropiado para entornos de producción con grandes volúmenes de datos, como es el caso del Censo Nacional.

7. Factores de Éxito y Riesgos

Entre los factores de éxito destaca el uso de métricas sólidas de evaluación y el ajuste fino de hiperparámetros, lo cual permitió obtener modelos competitivos. No obstante, existen riesgos potenciales, como la presencia de sesgos en los datos de entrada (por ejemplo, registros incompletos o errores de codificación) que pueden afectar la calidad de los clústeres. Además, la interpretación de los clústeres requiere conocimiento experto para evitar una mala asignación de etiquetas o decisiones basadas en agrupaciones poco significativas.

8. Impactos Potenciales

Una adecuada segmentación de hogares vulnerables puede tener un impacto directo en la eficiencia de las políticas públicas, ya que permite focalizar recursos en las zonas de mayor necesidad. En términos financieros, esto puede traducirse en ahorros al evitar intervenciones innecesarias o mal dirigidas, y maximizar el impacto de las inversiones sociales.

9. Informe Final (Técnico)

Durante la ejecución se observaron errores comunes como:
- Falsos positivos: hogares agrupados como vulnerables que no presentaban condiciones críticas reales.
- Falsos negativos: viviendas precarias que fueron agrupadas con hogares en mejor estado, diluyendo su riesgo.

Esto puede deberse a la falta de variables más específicas, como condiciones de acceso a servicios o ubicación geográfica exacta. Se recomienda incorporar más features, particularmente relacionadas con calidad de servicios públicos, riesgo geológico o exposición a desastres naturales. Además, explorar arquitecturas híbridas como modelos de clustering supervisado o reducción de dimensionalidad previa con técnicas como t-SNE o PCA podría mejorar la separación entre grupos.

7. Plan de Implementación (Deployment – CRISP-DM)

Para facilitar el uso práctico de los resultados del modelo de clustering, se propone una arquitectura sencilla pero robusta. Esta solución incluye un *pipeline* automatizado que se encargará de preparar y actualizar periódicamente los datos, incorporando información del Censo 2022 u otras bases administrativas relevantes proporcionadas por la Escuela Politécnica Nacional o el Instituto Geofísico.

El modelo final se implementaría en un entorno de producción a través de una API (interfaz de programación de aplicaciones), desarrollada con herramientas como FastAPI o Flask, que permiten compartir los resultados de forma rápida y segura. Esta API estaría alojada en un servidor con capacidad de procesamiento eficiente —por ejemplo, en la nube, utilizando servicios como AWS o Google Cloud Platform.

Para los usuarios finales (como investigadores, tomadores de decisiones o entidades públicas), se desarrollaría un dashboard interactivo con herramientas como Dash o Streamlit. Este dashboard permitiría visualizar de manera clara y dinámica los distintos grupos de vulnerabilidad detectados por el modelo, filtrando por provincia, cantón o zonas específicas.

Estrategia de Monitoreo y Mantenimiento

Se establecerán mecanismos de monitoreo automático del rendimiento del modelo a través de métricas como la estabilidad de clústeres y el cambio en la distribución de los datos. Se recomienda implementar un plan de reentrenamiento semestral o anual, dependiendo de la disponibilidad de nuevos datos censales o administrativos. Las herramientas de monitoreo pueden incluir MLflow, Prometheus y alertas con integración en plataformas como Slack o correo institucional.

Informe Final Técnico – Infraestructura y DevOps

El despliegue se planifica con entornos separados para desarrollo, pruebas y producción. Las herramientas DevOps sugeridas incluyen:
- Docker para la contenedorización del modelo.
- GitHub Actions o Jenkins para la automatización del despliegue.
- MLflow para el seguimiento de versiones del modelo.
- Terraform o CloudFormation para la infraestructura como código.

El plan de reentrenamiento debe considerar tanto cambios en la calidad de los datos como en las políticas públicas relevantes. Puede programarse con tareas periódicas (cron jobs) o con disparadores basados en cambios detectados en los datos.

Conclusiones, Próximos Pasos y Recomendaciones

Este proyecto logró construir y comparar dos modelos de clustering eficientes para la segmentación de hogares vulnerables, utilizando datos del Censo de Ecuador. KMeans mostró mejor rendimiento en todas las métricas, mientras que MiniBatchKMeans demostró mayor escalabilidad, lo cual es clave para su implementación futura.
Los próximos pasos incluyen:
- Incorporar variables adicionales como acceso a servicios básicos o riesgo geográfico.
- Aplicar modelos de clustering supervisado o semi-supervisado.
- Desarrollar mapas interactivos con visualización geoespacial.
- Implementar el modelo como herramienta de apoyo en instituciones públicas o ONGs.

Se recomienda priorizar la integración con otras bases de datos y establecer alianzas con entidades territoriales para mejorar la interpretación de los resultados. La documentación completa del pipeline

y del modelo está consolidada y lista para su transferencia a equipos de desarrollo o implementación.

ANEXOS:

	NOMBRE DE LA VARIABLE	CATEGORÍAS (Código y etiqueta)
I01	Provincia	De acuerdo a Clasificador Geográfico Estadístico
I02	Identificador de Cantón	De acuerdo a cartografía censal
I10	Número de vivienda	De acuerdo a cartografía censal
D01	Tipo de vía	1. Calle 2. Avenida 3. Carretera 4. Pasaje 5. Callejón 6. Sendero 7. Camino 8. Otro
V01	Tipo de vivienda	1. Casa/villa 2. Departamento en casa o edificio 3. Cuarto/s en casa de inquilinato. 4. Mediagua 5. Rancho 6. Covacha 7. Choza 8. Otra vivienda particular 9. Hotel, pensión, residencial u hostal 10. Cuartel militar, policía o bomberos 11 Centro de privación de libertad/cárcel 12.Hospital, clínica, etc. 13. Convento o institución religiosa 14. Centro de acogida y protección para niñas/os y

		adolescentes 15. Residencia de adultos mayores/Asilo de ancianos 16. Internado de estudiantes 17. Campamento de trabajo 18. Otra vivienda colectiva 19. Sin vivienda
V0201	Condición de ocupación de vivienda particular	1. Ocupada con personas presentes 2. Ocupada con personas ausentes 3. De temporada o vacacional 4. Desocupada 5. En construcción
V0202	Condición de ocupación de vivienda colectiva	1. Con residentes habituales 2. Sin residentes habituales
V03	Material predominante del techo o cubierta	1. Hormigón (losa, cemento)? 2. Fibrocemento, asbesto (eternit, eurolit)? 3. Zinc, aluminio (lámina o plancha metálica)? 4.Teja? 5. Palma, paja u hoja? 6. Otro material?
V04	Estado del techo o cubierta	1. Bueno? 2. Regular? 3. Malo?
V05	Material predominante de las paredes exteriores	1. Hormigón? 2. Ladrillo o bloque? 3. Panel prefabricado (yeso, fibrocemento, etc.)? 4. Adobe o tapia? 5. Madera?

		6. Caña revestida o bahareque? 7. Caña no revestida? 8. Otro material?
V06	Estado de las paredes exteriores	1. Bueno? 2. Regular? 3. Malo?
V07	Material predominante del piso	1. Duela, parquet, tablón o piso flotante? 2. Cerámica, baldosa, vinil o porcelanato? 3. Mármol o marmetón? 4. Ladrillo o cemento? 5. Tabla sin tratar? 6. Caña sin tratar? 7. Tierra? 8. Otro material?
V08	Estado del piso	1. Bueno? 2. Regular? 3. Malo?
V09	El agua que recibe la vivienda es	1. Por tubería, dentro de la vivienda? 2. Por tubería, fuera de la vivienda pero dentro del edificio, lote o terreno? 3. Por tubería, fuera del edificio, lote o terreno? 4. No recibe agua por tubería, sino por otros medios?
V10	El agua que recibe la vivienda proviene o es suministrada por	1. Empresa Pública/Municipio? 2. Juntas de Agua/Organizaciones comunitarias/GAD parroquial? 3. Pozo? 4. Carro o tanquero repartidor? 5. Otras fuentes (río, vertiente, acequia, canal, grieta o agua lluvia)?

V11	El servicio higiénico de la vivienda es	1. Inodoro o escusado, conectado a red pública de alcantarillado? 2. Inodoro o escusado, conectado a pozo séptico? 3. Inodoro o escusado, conectado a biodigestor? 4. Inodoro o escusado, conectado a pozo ciego? 5. Inodoro o escusado, con descarga directa al mar río, lago o quebrada? 6. Letrina? 7. No tiene
V12	Disponibilidad de energía eléctrica por red pública	1. Sí 2. No
V13	Disponibilidad de otra fuente de energía eléctrica	1. Planta eléctrica (generador de luz)? 2. Energía solar (panel fotovoltaico)? 3. Energía eólica (a partir del viento)? 4. Otra fuente (desechos vegetales y animales)? 5. No dispone
V14	Eliminación de la basura	1. Por carro recolector? 2. Por contenedor municipal? 3. La arroja en terreno baldío? 4. La quema? 5. La entierra? 6. La arroja al río, acequia, canal o quebrada? 7. De otra forma?
V15	Número de cuartos	1-20
V16	Todas las personas comparten un mismo gasto para la alimentación	1. Sí 2. No
V17	Número de hogares	1-9

AUR	Área urbana o rural	1. Área Urbana 2. Área Rural
CANTON	Cantón	De acuerdo a Clasificador Geográfico Estadístico
ID_VIV	Identificador de la vivienda	
TOTFALL	Total de fallecidos de la vivienda	0-99
TOTEMI	Total de emigrantes de la vivienda	0-99
TOTPER	Total de personas de la vivienda	0-9999
V0201R	Condición de ocupación de vivienda particular (recodificada)	1. Ocupada 2. De temporada o vacacional 3. Desocupada 4. En construcción
V15R	Número de cuartos (recodificada)	1. Un cuarto 2. Dos cuartos 3. Tres cuartos 4. Cuatro cuartos 5. Cinco cuartos 6. Seis o más cuartos
DEF_HAB	Déficit habitacional	1. Dignas o aceptables 2. Déficit cualitativo (Recuperables) 3. Déficit cuantitativo (Irrecuperables)
IMP_VOPA	Registro imputado en vivienda ocupada con personas ausentes	1. Sí 2. No

DATASET: HOGAR

CÓDIGO DE VARIABLE	NOMBRE DE LA VARIABLE	CATEGORÍAS (Código y etiqueta)
I01	Provincia	De acuerdo a Clasificador Geográfico Estadístico
I02	Identificador de Cantón	De acuerdo a cartografía censal
I10	Número de vivienda	De acuerdo a cartografía censal
INH	Número de hogar	00-10
H01	Número de dormitorios	0-20
H02	Cuarto o espacio exclusivo para cocinar	1. Sí 2. No
H03	Disponibilidad de servicio higiénico, inodoro o escusado	1. De uso exclusivo del hogar 2. Compartido con varios hogares 3. No tiene
H04	Disponibilidad de espacio con instalaciones y/o ducha para bañarse	1. De uso exclusivo del hogar 2. Compartido con varios hogares 3. No tiene
H05	Principal combustible o energía para cocinar	1. Gas de tanque o cilindro 2. Gas centralizado (por tubería) 3. Electricidad 4. Leña o carbón 5. Biogás (residuos vegetales y/o animales, etc.) 6. Otro (Ej: gasolina, kerex, diésel, etc.) 7. Ninguno (no cocina)

H06	Principalmente, el agua que beben los miembros del hogar	1. La beben, tal como llega al hogar? 2. La compran (agua envasada en bidón, botella o funda)? 3. La hierven? 4. Le ponen cloro? 5. La filtran (colocan filtros en el grifo o usan purificadores)? 6. Realizan otro tratamiento?
H0701	Acostumbra separar la basura en orgánica e inorgánica	1. Sí 2. No
H0702	Acostumbra separar desperdicios para dar a los animales o plantas	1. Sí 2. No
H0703	Acostumbra separar papel, cartón, plástico o vidrio para vender, regalar o reutilizar	1. Sí 2. No
H0801	Tiene este hogar perros	1. Sí 2. No
H0801N	Número de perros	1-98
H0802	Tiene este hogar gatos	1. Sí 2. No
H0802N	Número de gatos	1-98
H09	Tenencia de la vivienda	1. Propia y totalmente pagada? 2. Propia y la está pagando? 3. Propia (regalada, donada, heredada o por posesión)? 4. Arrendada/anticresis? 5. Prestada o cedida (no paga)? 6. Por servicios?

H1001	Dispone de servicio de teléfono convencional	1. Sí 2. No
H1002	Dispone de servicio de teléfono celular	1. Sí 2. No
H1003	Dispone de servicio de televisión pagada	1. Sí 2. No
H1004	Dispone de servicio de internet fijo	1. Sí 2. No
H1005	Dispone de computadora	1. Sí 2. No
H1006	Dispone de refrigeradora	1. Sí 2. No
H1007	Dispone de máquina lavadora de ropa	1. Sí 2. No
H1008	Dispone de máquina secadora de ropa	1. Sí 2. No
H1009	Dispone de horno microondas	1. Sí 2. No
H1010	Dispone de máquina extractora de olores	1. Sí 2. No
H1011	Dispone de automóvil o camioneta para uso exclusivo	1. Sí 2. No
H1012	Dispone de motocicleta para uso exclusivo	1. Sí 2. No
H11	Alguna persona falleció en los últimos tres años (a partir de enero 2020)	1. Sí 2. No 9. Se ignora
H1101	Número de personas fallecidas	1-20

H12	Alguna persona viajó a otro país y todavía no regresa (a partir de nov. 2010)	1. Sí 2. No 9. Se ignora
H1201	Número de personas emigrantes	1-20
H1301	Total de hombres en el hogar	0-9999
H1302	Total de mujeres en el hogar	0-9999
H1303	Total de personas en el hogar	1-9999
H15	Existe alguna persona que no haya sido mencionada en el hogar	1. Sí 2. No 9. Se ignora
AUR	Área urbana o rural	1. Área Urbana 2. Área Rural
CANTON	Cantón	De acuerdo a Clasificador Geográfico Estadístico
ID_VIV	Identificador de la vivienda	
ID_HOG	Identificador del hogar	
H01R	Número de dormitorios (recodificada)	0. Ninguno 1. Un dormitorio 2. Dos dormitorios 3. Tres dormitorios 4. Cuatro dormitorios 5. Cinco dormitorios 6. Seis o más dormitorios

IMP_VOPA	Registro imputado en vivienda ocupada con personas ausentes	1. Sí 2. No