Monitoreo del nivel de prevención ante sismos por distritos del Perú*

*Profesora: Soledad Espezúa

Amy Antalu Checcllo Huachua
Ingeniería de la Información
Universidad del Pacífico
Lima, Perú
aa.checcllloh@alum.up.edu.pe

Miguel Alexis Muñoz Chavez
Ingeniería de la Información
Universidad del Pacífico
Lima, Perú
ma.munozc@alum.up.edu.pe

Alexia Shariann Ríos Alarcón
Ingeniería de la Información
Universidad del Pacífico
Lima, Perú
as.riosa@alum.up.edu.pe

Alexandra Constanza Sanjinez Mendoza
Ingeniería de la Información
Universidad del Pacífico
Lima, Perú
ac.sanjinezm@alum.up.edu.pe

***Abstract*—Los sismos se caracterizan por ser impredecibles y destructivos; por ello, se busca implementar medidas preventivas para poder apaciguar las consecuencias negativas pero, en ocasiones, no se realizan adecuadamente. En el caso de Perú, se encuentra situado en una región geológicamente activa; sin embargo, está en el puesto 13 del ranking de naciones más vulnerables ante desastres naturales. Por ello, se busca disminuir el puesto del Perú en el ranking antes mencionado. Para ello, se unificaron diferentes conjuntos de datos para obtener una data centralizada con datos relevantes para analizar el nivel de prevención de distritos del Perú con técnicas de Data Mining como clustering. Se logró identificar las zonas donde el Perú se encuentra menos preparado para afrontar un sismo.**

***Index Terms*—sismos, prevención, clustering, suelos, distritos**

I. Introducción

Los sismos son eventos naturales que representan la liberación de energía acumulada en el interior de la Tierra, manifestándose en forma de vibraciones que pueden alcanzar diferentes magnitudes (Instituto Nacional de Defensa Civil, s.f.). Estos fenómenos geológicos, a menudo impredecibles y destructivos, requieren una atención constante en términos de prevención y preparación para minimizar sus impactos.

Para la elaboración del presente trabajo, se tomaron como referencia 2 informes técnicos. El primero fue realizado por la autoridad de CENEPRED con la colaboración de otras instituciones, como el Instituto Geofísico del Perú (IGP). En este informe, se evaluó el impacto de la ocurrencia de sismos de 8.8 Mw. Además de datos geológicos, se consideró un mapa de los establecimientos de salud [11]. En este estudio, se identificaron inicialmente los puntos geográficos más vulnerables a desastres naturales y, al combinarlos con el mapa de establecimientos de salud, se simularon escenarios para determinar su nivel de riesgo. Se observó un alto riesgo en las áreas cercanas a la costa del Perú, con un nivel de riesgo muy alto del 44.6%. Las provincias con mayor concentración de población en esta categoría fueron Lima y Callao.

En un segundo informe, llevado a cabo por el Banco Mundial y el estado peruano, se buscó evaluar el impacto de los sismos en la infraestructura educativa y, a partir de estos datos, proponer estrategias para enfrentar esta situación. Se presentaron datos, como la distribución de instituciones educativas, mapas probabilísticos de ocurrencia de sismos, mapas de riesgos sísmicos y mapas de tipos de suelo, entre otros [12]. Como conclusión, se determinó que el 51% de las edificaciones tienen un alto riesgo de colapso, mientras que el 21% presenta un daño potencial. A partir de estos resultados, el estudio propuso programas correspondientes de sustitución, refuerzo y medidas de intervención contingente como estrategias para reducir el riesgo sísmico.

El Perú, un país situado en una región geológicamente activa, se encuentra en el puesto 13 del ranking de naciones más vulnerables ante desastres naturales. Esta vulnerabilidad no solo se limita a los sismos, sino que abarca una serie de amenazas naturales, como inundaciones, deslizamientos de tierra y eventos climáticos extremos. Esta situación demanda una respuesta efectiva y un enfoque proactivo en la prevención y mitigación de estos eventos para salvaguardar vidas y bienes.

Nuestro objetivo principal es la disminución del puesto de Perú en el ranking de países más vulnerables ante desastres naturales. Para lograr este cometido, es imperativo no sólo reaccionar ante las consecuencias de estos eventos, sino también adoptar un enfoque integral que priorice la prevención y la preparación en todos los niveles de gobierno y en cada rincón del país. La reducción de la vulnerabilidad y el fortalecimiento de la resiliencia son fundamentales para asegurar un futuro más seguro y sostenible para la población peruana.

TABLE I
DESCRIPCIÓN DE LOS DATOS DEL IGP

Variable	*Descripción*
Fecha UTC	Fecha universal del sismo ocurrido.
Hora UTC	Hora universal del sismo ocurrido.
Latitud	Latitud del epicentro del sismo.
Longitud	Longitud del epicentro del sismo.
Profundidad	Profundidad (en kilómetros) del epicentro del sismo.
Magnitud	Magnitud del sismmo en escala de Richter.

[a]Elaboración propia.

TABLE II
DESCRIPCIÓN DE LOS DATOS DE LA ENAHO

Variable	*Descripción*
Year	Año realizado la encuesta.
DOMINIO	Dominio geográfico de la zona de la vivienda: 1(Costa Norte), 2(Costa Centro), 3(Costa Sur), 4(Sierra Norte), 5(Sierra Centro), 6(Sierra Sur), 7(Selva), 8(Lima Metropolitana).
$\text{Tipo}_{vivienda}$	1(Casa independiente), 2(Departamento en edificio), 3(Vivienda en quinta), 4(Vivienda en casa de vecindad), 5(Choza), 6(Vivienda improvisada), 7(Local no destinado para habitación humana).
$\text{Pared}_{material}$	Material predominante de las paredes: 1(adobe), 2(piedra, sillar), 3(adobe), 4(tapia), 5(Quincha (caña con Barro), 6(Piedra con barro), 7(Madera), 8(Triplay), 9(Otro).
$\text{Piso}_{material}$	Material predominante de los pisos: 1(madera pulida), 2(láminas asfálticas), 3(losetas), 4(madera), 5(Cemento), 6(Tierra), 7(Otro).
$\text{Techo}_{material}$	Material predominante de los techos: 1(concreto armado), 2(madera), 3(tejas), 4(calamina), 5(Triplay), 6(Paja), 7(Otro).
Licencia	La vivienda cuenta con licencia de construcción: 1(Si), 2(No), 3(No sabe)
$\text{Asistencia}_{tecnica}$	Vivienda construida con asistencia técnica de Ing.Civil o arquitecto: 1(Si), 2(No), 3(No sabe)
$\text{Nivel}_{vegetedad,vivienda}$	La vivienda hasta qué nivel es de su propiedad 1(alquilada), 2(propi, pago completo), 3(propia, invasión), 4(propia, a crédito), 5(cedida por trabajo), 6(cedida), 7(Otra)
$\text{Titulo}_{propiedad}$	Tiene título de propiedad: 1(Si), 2(No), 3(En trámite)
Registro_{UNAHP}	Vivienda registrada en la SUNARP: 1(Si), 2(No)
$\text{Techo}_{material}$	Material predominante de los techos: 1(concreto armado), 2(madera), 3(tejas), 4(calamina), 5(Triplay), 6(Paja), 7(Otro)
NBI1	Necesidad Básica Insatisfecha 1: Vivienda adecuada 0(Si), 1(No)
NBI2	Necesidad Básica Insatisfecha 2: Vivienda sin hacinamiento 0(Si), 1(No)
NBI3	Necesidad Básica Insatisfecha 3: Vivienda con SSHH 0(Si), 1(No)
NBI4	Necesidad Básica Insatisfecha 4: Vivienda con niños que asisten a la escuela 0(Si), 1(No)
NBI5	Necesidad Básica Insatisfecha 5: Vivienda sin dependencia económica 0(Si), 1(No)

[a]Elaboración propia.

La falta de capacidades por parte del Estado y su mala capacidad de adaptación a la realidad geológica y climática del Perú son los principales obstáculos para alcanzar la situación deseada. La falta de inversión en infraestructuras resistentes a sismos, la limitada educación en prevención de desastres y la falta de coordinación entre las diferentes instituciones gubernamentales son algunas de las deficiencias que debemos abordar. Estos desafíos son cruciales para garantizar la seguridad y el bienestar de la población peruana

II. RECOPILACIÓN Y PREPARACIÓN DE DATOS

Luego de decidir el tema, se procede a recopilar las fuentes de datos necesarias para poder realizar el análisis correspondiente. Se utilizaron cuatro fuentes de datos diferentes.

A. Instituto geofísico del Perú (IGP)

Para el presente trabajo, se procedió a extraer los datos de los sismos del portal del IGP en donde se accede a la página de Descargar datos y se procede a descargar la data del periodo 2004 - 2023 (hasta el mes de agosto) y se mantienen las demás opciones predeterminadas de la página. Si bien es cierto que se ha descargado la data hasta agosto de 2023 sólo se utilizarán los registros hasta el año 2022.

El archivo descargado se encuentra en formato de excel en la cual se encuentran los siguientes atributos: fecha UTC, hora UTC, latitud, longitud, profundidad, magnitud. En la Tabla I, se muestran las descripciones de cada atributo.

Después de una breve descripción de los atributos de la base de datos del IGP, se procede a explicar la limpieza y preprocesamiento de estos datos que luego se unirán a una data centralizada.

Primero, se realiza una exploración general del dataset para identificar posibles incongruencias o datos nulos. En un primer alcance, no se presentan datos nulos y, el dataset tiene 12880 registros y 6 atributos. Sin embargo, la data debe adecuarse y transformarse de manera que brinde valor a la data centralizada. Por ello, los atributos latitud y longitud, que usualmente son eliminados debido a que no brindan mayor información o valor, deben ser transformados en información interpretable como direcciones.

Segundo, teniendo en cuenta lo anterior, se transforman los datos de latitud y longitud en 4 nuevos atributos los cuales son país, región, provincia, distrito. GeoPy facilita a desarrolladores de Python localizar las coordenadas de direcciones, ciudades, países y puntos de referencia en todo el mundo mediante geocodificadores de terceros y otras fuentes de datos [2].

Tercero, luego de la conversión de los atributos latitud y longitud las dimensiones del conjunto de datos cambia y se vuelve a explorar. Por un lado, los atributos pasan de ser 6 a 10 y se presentan datos nulos en 3 de las 4 columnas agregadas: Región, Provincia, Distrito. En el caso del atributo Región, se logra discernir que los datos faltantes correspondes a sismos que ocurrieron en el mar. Si bien es cierto, se puede etiquetar estos datos con el nombre "Mar" no brindaría mucho valor al análisis de las viviendas por distrito. Por ello, se eliminan aquellos registros. Luego, se realiza la revisión de datos nulos en los atributos de Provincia y Distrito. Para decidir qué hacer con estos datos, se evalúa si los registros corresponden a sismos con epicentro en Perú y no en otro país. Una vez realizada la revisión, se observa que los datos nulos corresponden a otros países; por ello, también se decide eliminarlos. Finalmente, se observan los datos nulos del atributo Distrito. En este caso, se realizaron dos acciones: eliminar los registros no nulos que corresponden a otros país o que no se lograron identificar e imputar los datos nulos de los distritos con el nombre de la provincia. Para el caso de la imputación, también se realiza la actualización de la latitud y la longitud de acuerdo a la combinación de región, provincia y distrito.

Cuarto, una vez realizado lo anterior, se obtiene un dataframe con 7697 registros y 10 atributos. De esta manera, se procede a crear la "key" que va a permitir que este conjunto de datos se una a la data centralizada. Esta clave se conforma por el nombre de la región, provincia, distrito y año. Cabe mencionar que, una vez obtenido el atributo año, se eliminan los registros que corresponden al 2023.

Finalmente, se crea un nuevo dataset a partir de las estadísticas del conjunto de datos que ha sido tratado. Se crean las siguientes columnas: conteo de sismos, promedio de sismos, magnitud mínima de sismos, magnitud máxima de sismos.

B. ENAHO

La descarga de datos se obtiene del portal del Instituto Nacional de Estadística e Informática desde el año 2004 hasta el 2022. De manera complementaria, se obtienen datos de ubigeo para poder identificar la región, provincia y distrito de las viviendas encuestadas de la base de datos de la ENAHO.

En lo que respecta la limpieza y el preprocesamiento de los datos, se realiza la descarga de los datos de todos los años, para luego realizar su posterior concatenación. Teniendo en cuenta

TABLE III
DESCRIPCIÓN DE LOS DATOS DEL MINSA

Variable	***Descripción***
Nombre del establecimiento	Nombre del establecimiento
Clasificación	Clasificación del establecimiento
Tipo	Establecimiento con o sin internamiento
Departamento	Nombre del departamento
Provincia	Nombre de la provincia
Distrito	Nombre del distrito
Dirección	Dirección del establecimiento
Categoría	Categoría del establecimiento (I1,I2,I3 o I4)
Teléfono	Número de teléfono de contacto
Horario	Horario de funcionamiento
Año	Año de inicio de actividades

[a]Elaboración propia.

que los datasets corresponden a diferentes años, el problema que se presenta es la diferencia en algunas preguntas, por eso se dividen en dos siendo uno el dataframe correspondiente a los años 2004 - 2011 que no contienen las variables: Licencia, Asistencia tecnica y Registro SUNARP; por otro lado, el dataframe correspondiente a los años 2012 - 2022 si contiene las variables: Licencia, Asistencia tecnica y Registro SUNARP.

C. Ministerio de Salud (MINSA)

Se descargaron los datos actualizados de establecimientos de salud de primer nivel de atención en el Perú hasta el 25 de enero de 2023, proporcionados por el Ministerio de Salud. El archivo descargado se presenta en formato Excel y consta de las siguientes columnas: Nombre del establecimiento, clasificación, tipo, departamento, provincia, distrito, dirección, categoría, teléfono y horario. En total, se registraron 8,729 establecimientos de salud. Además, se incluyó una columna adicional, denominada "AÑO", la cual representa el año en que cada centro de salud comenzó sus operaciones. Esta columna se agregó ya que se utilizará en la concatenación con un conjunto de datos más amplio. Más adelante, se proporcionarán detalles sobre la obtención de los años de inicio de actividades.

En la siguiente tabla, se muestra todas las variables junto a su descripción:

Es importante definir la variable categoría porque es la de mayor interés para el análisis de prevención. Existen 4 tipos de categorías: la categoría I-1 son los establecimientos de salud que cuentan con profesionales de la salud, pero no tienen médicos cirujanos, la categoría I-2 son los puestos de salud o posta de salud (con médico). Además de los consultorios médicos (con médicos con o sin especialidad), la categoría I-3 corresponde a los centros de salud, centros médicos, centros médicos especializados y policlínicos y la categoría I-4 agrupa a los centros de salud y los centros médicos con camas de internamiento. La preparación de los datos incluyó la adición de una columna que reflejara el año de inicio de actividades de los establecimientos de salud. Para llevar a cabo este proceso, se empleó la técnica de web scraping, accediendo a la siguiente página web del Ministerio de Salud: https://www.establecimientosdesalud.info/. Una vez completado el dataset, se procedió a realizar dos ajustes importantes. Primero, se reemplazaron todos los años anteriores a 2004. Además, se corrigieron los 40 valores faltantes, que representaban un 0.5% del total de registros, asignándoles el año 2004, ya que este corresponde al año mínimo común con los demás conjuntos de datos utilizados. Posteriormente, se llevó a cabo la conversión de la variable "categoría" en variables dummy, lo que resultó en la creación de columnas con valores de ceros y unos para cada tipo de categoría (I1, I2, I3 e I4).

A continuación, se generó una tabla con los valores únicos de Departamento, Provincia y Distrito, junto con los años desde 2004 hasta 2022. Esta tabla tiene un total de 35,424 filas, resultado de la combinación de 1866 valores únicos de Departamento, Provincia y Distrito durante un período de 19 años. Luego, se diseño una función para calcular la suma de la cantidad de establecimientos en cada categoría a lo largo de los años. Esto permitió obtener la cantidad acumulada de establecimientos por categoría en el año 2022. Para lograr esto, se modificaron los años anteriores a 2004, asignándoles el valor 2004, lo que permitió incluirlos en el cálculo acumulado de dicho año. Finalmente, se creó una variable "key" que representa la combinación de Departamento, Provincia, Distrito y Año. Esta variable se utilizó para facilitar la integración con otros conjuntos de datos. Las variables que contribuyeron a crear esta clave se eliminaron, y lo que permanece en la tabla son las sumatorias totales por categoría y key.

D. Tipos de suelo

Para la obtención de las etiquetas del tipo de suelo presentes en los distritos del Perú se utilizó la información pública de la siguiente página web: www.geogpsperu.com. Utilizamos el programa QGIS, que es un Sistema de Información Geográfica (SIG). En este programa, abrimos los archivos en formato shapefile (de los suelos) y los interceptamos con el archivo de límite distrital, lo que nos permite obtener el resultado deseado nivel de distrito con los tipos de clasificación de suelos, el resultado es el que se ve en la tabla IV.

TABLE IV
DESCRIPCIÓN DE LA BASE DE DATOS DE SUELOS

Variable	**Descripción**
Capital	Variable nominal que indica nombre de la capital
NombreDep	Variable nominal que indica nombre del departamento
NombreProv	Variable nominal que indica nombre de la provincia
NombreDist	Variable nominal que indica nombre del distrito
Ubigeo	Identificador único de la ubicación geográfica
KM2	Dimensiones en Km2
SIMSUE	Acrónimo según el tipo de clasificación
Descripción	Clasificación del tipo de suelo
Proporción	Proporción de los tipos de suelo
Paisaje	Descripción del paisaje
Pendiente	Pendiente del terreno de inclinación
Longitud	Medida angular que indica la ubicación este-oeste de un punto en la superficie
Latitud	Distancia angular medida en grados, minutos y segundos al norte o al sur del ecuador

En un principio, el archivo tenía unas dimensiones de 2048 filas y 12 columnas, que representan los atributos detallados

en la tabla X. No obstante, con el propósito de integrarlo con otras bases de datos, se agregó una columna adicional que funciona como clave primaria. Esta clave primaria se compone del nombre del departamento, nombre de la provincia, nombre del distrito y el año. En este caso, se asumió que la clasificación del suelo permanece constante durante el período de 2004 a 2022. Sin embargo, es importante señalar que esta suposición no siempre es precisa, ya que eventos naturales como huaycos, aluviones e inundaciones pueden modificar la topografía del terreno. La fuente de datos utilizada para extraer esta información tiene sus limitaciones. Para investigaciones futuras, sería beneficioso identificar y tener en cuenta estos cambios para mejorar el análisis. Finalmente, se obtuvo una base de datos con dimensiones de 38912 x 13.

A continuación se tiene una descripción breve de los términos utilizados para la descripción de los suelos, las descripciones se obtubieron de la pagina de la The Food and Agriculture Organization (FAO):

1) Leptosol éutrico: Un Leptosol con buenas propiedades nutricionales.
2) Regosol éutrico: Un Regosol con buenas propiedades nutricionales.
3) Afloramiento lítico: Suelo con rocas expuestas en la superficie.
4) Kastanozem háplico: Un tipo de suelo rico en nutrientes con un horizonte oscuro.
5) Regosol dístrico: Un Regosol con características particulares de drenaje.
6) Cambisol éutrico: Un tipo de suelo con buenos nutrientes y característicasespecíficas.
7) Fluvisol éutrico: Suelo fluvial con buenas propiedades nutricionales.
8) Gleysol éutrico: Un Gleysol con buenas propiedades nutricionales.
9) Andosol úmbrico: Un tipo de suelo volcánico rico en nutrientes.
10) Calcisol háplico: Suelo con calcio y horizontes oscuros.
11) Vertisol éutrico: Un tipo de suelo rico en nutrientes con cara cterísticas específicas de expansión y contracción.
12) Andosol móllico: Suelo volcánico con horizontes oscuros.
13) Lixisol háplico: Suelo con características particulares de lixiviación.
14) Gleysol dístrico: Un Gleysol con características particulares de drenaje.
15) Fluvisol dístrico: Suelo fluvial con características específicas de drenaje.

III. Exploración de Datos

Despues de integrar los datos segun se muestra en la Figura 1 se continuo con la exploración y visualización de los datos. Previamente, se realiza un análisis por cada conjunto de datos antes de la data centralizada.

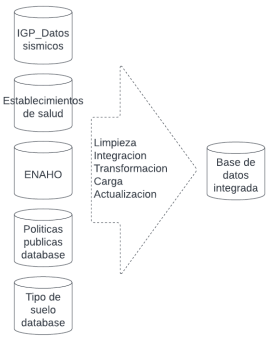

Fig. 1. Integracion de los dataset

A. Dataset Sismos

En la Fig. 2, se puede observar que a partir de los datos de la latitud y longitud se diseña una aproximación del mapa del Perú. En base a ello, se puede observar que la mayoría de los sismos se encuentran localizados en la zona costa y sierra del Perú.

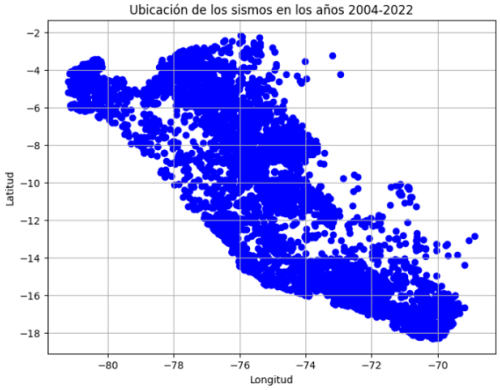

Fig. 2. Ubicación de los sismos en los años 2004-2022

En la Fig. 3, se puede visualizar que, por lo general, ocurren más de 100 sismos. No obstante, se puede observar que mayormente ocurren más de 250 sismos siendo pocos años los que ocurren menos de esa cantidad.

En la Fig. 4, se puede observar que el promedio de la magnitud de sismos por año no varían significativamente. Esto podría dar a entender que por lo general la magnitud de sismos se mantiene constante.

Teniendo en cuenta los anteriores gráficos, se decide realizar un análisis de los sismos al comienzo del periodo del estudio (2004) y al final (2022) para conocer cómo cambian los sismos y si se repiten o se concentran en algunas zonas teniendo.

En la Fig. 5, tal como se identifica en la Fig. 1, se logra discernir que los sismos se concentran en la zona costa y sierra del Perú. No obstante, se puede discernir que existe una mayor concentración de sismos en el norte.

Finalmente, en la Fig. 6, se puede observar que los sismos se concentran en la zona costa y sierra; no obstante, a difer-

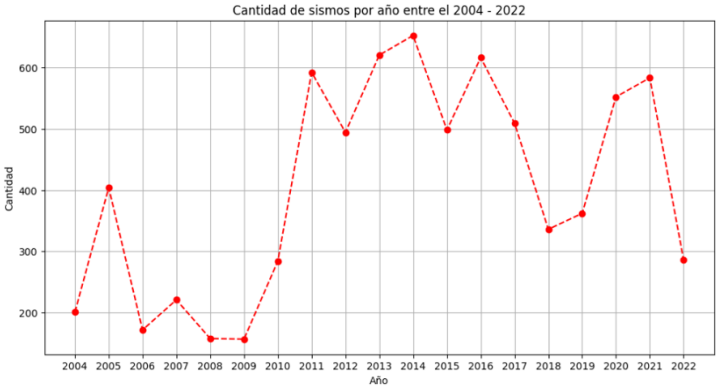

Fig. 3. Cantidad de sismos por año entre el 2004-2022

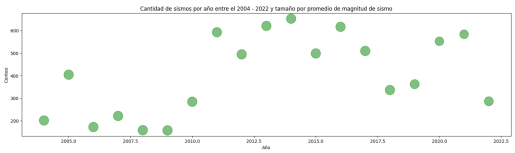

Fig. 4. Cantidad de sismos por año entre el 2004 - 2022 y tamaño por promedio de magnitud de sismo

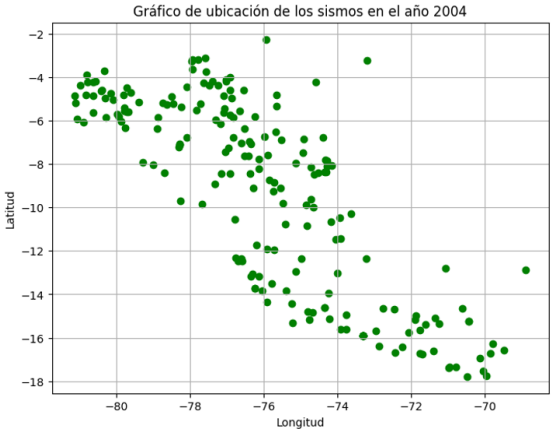

Fig. 5. Ubicación de los sismos en el 2004

encia del gráfico del 2004, los sismos presentan una mayor concentrancón en el sur y en el centro.

B. Dataset Establecimientos de salud

En la Fig.7, se puede observar que en el Perú la mayor cantidad de establecimientos de salud pertenecen a la categoría I-1, lo cual indica un bajo nivel de calidad de los establecimientos de salud. Asimismo, es importante este gráfico porque se pueden tener una cantidad alta de establecimientos de salud; sin embargo, eso no significa que todos sean de una alta calidad.

La siguiente Fig.8, es un gráfico de barras con los 10 departamentos que tienen la menor cantidad de establecimientos de salud. Este gráfico permite identificar los departamentos cuyos distritos podrían tener un menor nivel de prevención.

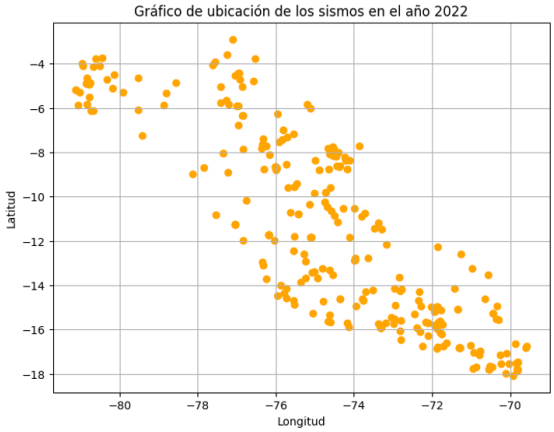

Fig. 6. Ubicación de los sismos en el 2022

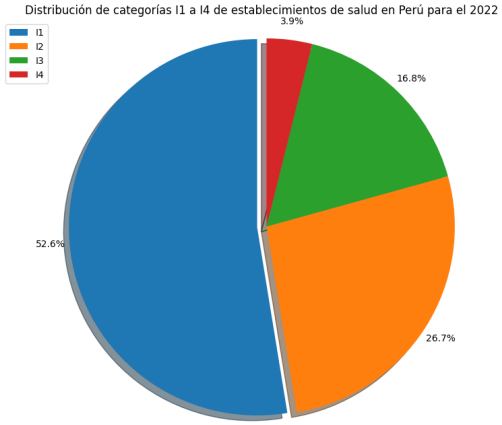

Fig. 7. Gráfico de pie chart de las categorías de los establecimientos

Se muestra que Tumbes es el departamento con una menor cantidad de establecimientos con 44.

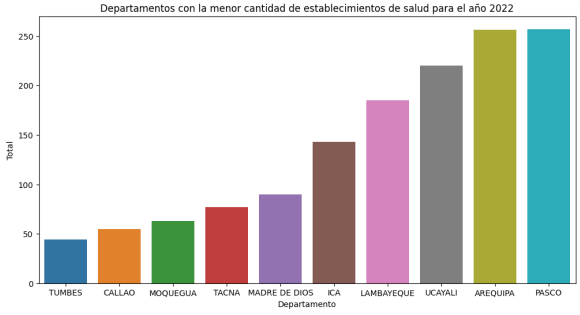

Fig. 8. Gráfico de barras de los 10 departamentos con la menor cantidad de establecimientos de salud para el año 2022

A continuación, la Fig.9 es un gráfico de barras con los 10 departamentos que tienen la mayor cantidad de establecimientos de salud. Este gráfico muestra que el departamento de Cajamarca tiene la mayor cantidad de establecimientos de salud con 863. A continuación, el departamento de Lima con 723 establecimientos.

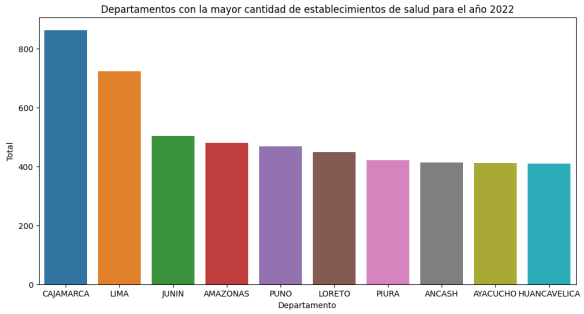

Fig. 9. Gráfico de barras de los 10 departamentos con la mayor cantidad de establecimientos de salud para el año 2022

Finalmente, en la Fig.10 se muestra un gráfico de líneas para ver en detalle como ha sido la evolución en la cantidad de establecimientos en el departamento con la mayor cantidad. Se puede observar que en el periodo de 2006 a 2010 se tiene el mayor aumento en la cantidad de establecimientos.

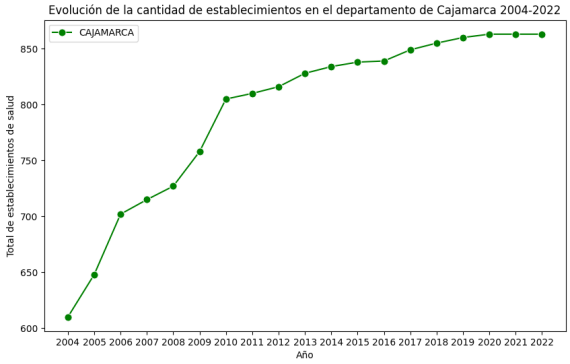

Fig. 10. Gráfico de líneas de la evolución de establecimientos de salud en Cajamarca

C. Dataset Suelos

Segun la Fig. 11 el suelo "Fluvisol éutrico - Gleysol éutrico" posee un (27.9%) es un tipo de suelo se encuentra en áreas con antecedentes de inundaciones o un alto contenido de agua estancada, lo que puede influir en su capacidad para resistir sismos. en Km2 tiene un valor de 69402.158

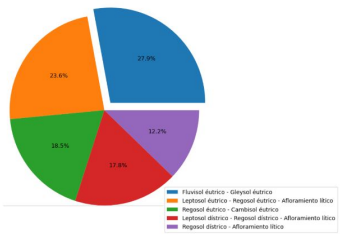

Fig. 11. Gráfico top 5 suelos con mas km2

D. Dataset ENAHO

Las regiones extraídas son las que cuentan con una mayor densidad poblacional y por ende son de mayor interés, pues su proporción de viviendas en situación de Asistencia o Calidad de materiales de construcción ocupan una mayor muestra de vivendas. En la Fig. 12 se muestra la proporción de viviendas que tienen paredes de ladrillo o cemento entre 2012 y 2022. Se muestra que Ica, Lima provincia y Lambayeque tiene la mayor proporción de viviendas con paredes de ladrillo.

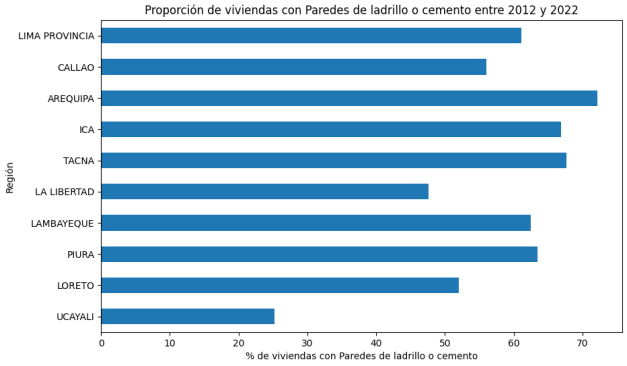

Fig. 12. Proporción de viviendas con Paredes de ladrillo o cemento entre 2012 y 2022

En la Fig. 13 se muestra la proporción de viviendas construídas con Asistencia Técnica entre 2012 y 2022. Se muestra que Ica, Lambayeque, Lima provincia y Arequipa son los departamentos con mayor proporción de viviendas con asistencia técnica. Lo cual indica que debe ser una vivienda de mayor calidad por la asistencia técnica. Se encuentra una

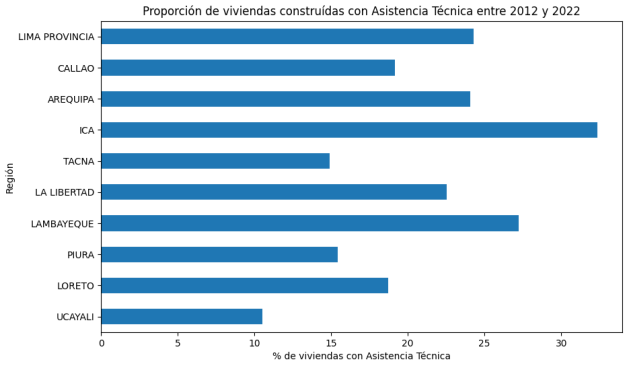

Fig. 13. Proporción de viviendas construídas con Asistencia Técnica entre 2012 y 2022

relación entre las viviendas construidas con asistencia técnica y las que tienen paredes construidas de ladrillo o cemento.

IV. Metodología

1) Data mining: La minería de datos es una técnica relativamente reciente para extraer conocimiento de vastas cantidades de datos. Esta práctica implica utilizar y procesar datos disponibles para tomar decisiones [8]. Implica la exploración de modelos en conjuntos de datos extensos utilizando técnicas que se sitúan en la intersección de aprendizaje automático, estadísticas y sistemas de bases de datos [9]. Facilita el análisis de patrones, como la categorización de datos a través de

estudios de agrupamiento, la detección de registros anómalos, también conocida como detección de anomalías, y las reglas o dependencias asociadas

2) Metodo KDD: Knowledge Discovery Dictionary (KDD): Este proceso abarca la recolección y el descubrimiento de datos e información, involucrando operaciones como el procesamiento, la selección y la preparación de datos, junto con la creación de información en conjuntos de datos y la interpretación de los enfoques más eficaces según los resultados observados [10]. Se caracteriza por una secuencia iterativa de integración de datos y la identificación de patrones en Minería de Datos (DM).

V. Aplicación de Técnicas de Data Mining

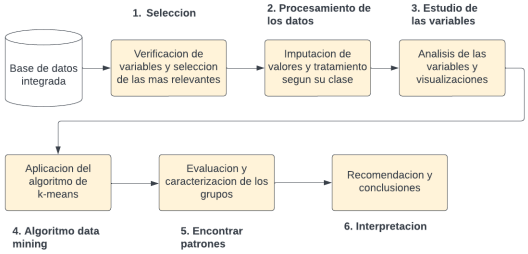

Fig. 14. Metodologia del estudio

Segun la estructura de la Fig. 14, comenzamos con la datos integrados de las diferentes fuentes seleccionadas, en un primer paso revisamos las variables y comenzamos la seleccion de las variables mas importantes. Luego, completamos valores faltantes de presentarse y dependiendo de la clase da datos se hacen transformacion ya sea "label encoding" o "one-hot encoding". Seguimos con la parte de visualizaciones para generar primeras observaciones y analizar las variables. Aplicamos el algoritmos de K-means para generar clusters y poder perfilar los grupos para posteriormente encontrar patrones interesantes que serviran para poder resolver las preguntas de investigacion.

VI. Resultados

A modo de resultados, se obtiene el análisis de todos los grupos de datos y la integración de todos en una sola data centralizada para su posterior análisis a través de clustering. Se selecciona un año de muestra con el propósito de llevar a cabo pruebas de agrupación en relación al nivel de prevención ante sismos de los distritos analizados. En este proceso se consideran los siguientes criterios:

- El año de muestra debe ser igual o posterior a 2017, ya que a partir de este año se cuenta con información disponible en todas las bases de datos requeridas para el análisis.
- El año de muestra debe ser aquel en el cual se registró la mayor cantidad de sismos. Esto se debe a que, al utilizar un año con un alto número de sismos, es posible examinar de manera más efectiva la frecuencia de ocurrencia de eventos sísmicos en áreas que presentan recurrencia en este tipo de fenómenos.
- Se ha decidido no incluir la Base de Datos de Tipo de Suelo en este análisis. La razón detrás de esta decisión es que, a pesar de contar con información sobre los tipos de suelo en la región, no se dispone de una referencia exacta que vincule estos tipos de suelo con el nivel de vulnerabilidad de las viviendas. Esta información es esencial para lograr una interpretación más detallada y precisa en el contexto de la prevención ante sismos.

A continuación se observa la imagen de la Silloutte Score para obtener la cantidad de clusters adecuados, se decide por k=3.

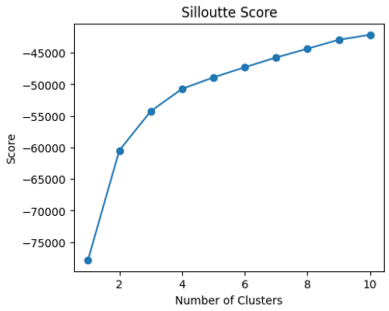

Fig. 15. Silloutte score

Luego en el entrenamiento del modelo de Clustering se agrupa los resultados obtenidos en relación dell clúster asignado para analizar la media de los valores de las variables para cada uno de los tre clústers asignados. En la Fig. 16, se evidencia que el procedimiento de agrupación ha conducido a la asignación de las filas con valores más elevados en todas sus variables al Cluster 1, seguido por el Cluster 2, y finalmente al Cluster 0. No obstante, este resultado no proporciona una agrupación ni una interpretación satisfactorias.

cluster_std	0	1	2
DOMINIO	4.887423	4.000000	5.169811
ESTRATO	6.160950	2.000000	2.377358
Registro_SUNARP_1	3.107300	148.750000	41.245283
Registro_SUNARP_2	0.966579	18.500000	7.169811
Asistencia_Tecnica_1	1.123131	90.500000	24.084906
Asistencia_Tecnica_2	13.290237	180.333333	64.896226
Asistencia_Tecnica_3	0.797713	52.416667	13.641509
Licencia_1	1.568162	98.083333	29.405660
Licencia_2	12.729112	172.916667	59.311321
Licencia_3	0.913808	52.250000	13.905660
Tipo_Vivienda_1	15.294635	296.083333	90.547170
Tipo_Vivienda_2	0.197010	28.583333	9.811321
Tipo_Vivienda_3	0.036939	4.250000	1.500000
Tipo_Vivienda_4	0.435356	8.083333	5.113208
Tipo_Vivienda_5	0.269129	0.000000	0.141509
Tipo_Vivienda_6	0.004398	0.500000	0.037736
Tipo_Vivienda_7	0.000000	0.000000	0.009434
Tipo_Vivienda_8	0.000000	0.000000	0.000000

Fig. 16. Tabla de agrupación por cada clúster

Por último, como se observa en la Fig. 17, la Clusterización no refleja un adecuado agrupamiento de los datos. Sin embargo, una de las hipótesis extraídas respecto a este resultado

es que se ha trabajado en su mayoria con variables binarias, las cuales no permiten al kmeans generar un óptimo modelo. Esto porque el kmeans es un algoritmo basado en distancias, asi que si se tiene variables con rangos más pequeños que otros, a pesar de la estandarización, no ayuda lo sufiente. Para una continuación del proyecto, se intentará codificar manualmente las variables categóricas con respecto al impacto que tendría en el Nivel de prevención ante sismos del Distrito, sería un análisis más exhaustivo de cada una de las variables, pero que se recompensaría enormemente a un mejor resultado e interpretación.

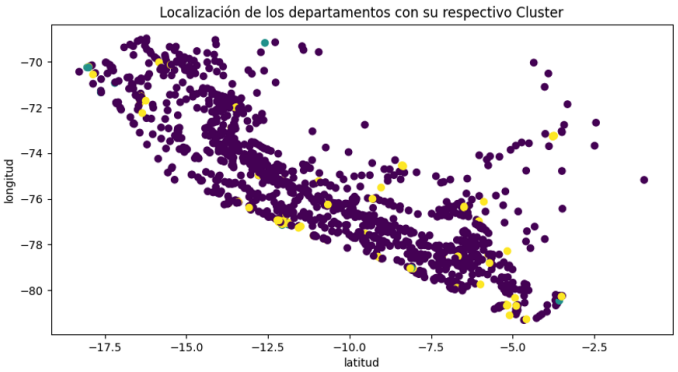

Fig. 17. Localización de los departamentos con su respectivo cluster

VII. Conclusiones

En conclusión, las variables de infraestructura de vivienda, tipo de suelo, actividad sísmica, disponibilidad de establecimientos de salud y políticas públicas influyen en la detección de patrones que reflejan el nivel de preparación ante sismos en los distritos del Perú.

Al analizar por partes, se logra discernir las zonas que suelen ser más afectadas por los sismos. Principalmente, se observa que la zona costa y sierra; pero a nivel más específico se identifica que usualmente los sismos se concentran en la zona costa norte, costa sur y sierra centro.

En los establecimientos de salud, se ha logrado identificar que la mayoría de establecimientos se agrupan en la categoría I-1 que hace referencia a establecimientos que cuentan con profesionales de la salud, pero no tienen médicos cirujanos. Es el nivel más bajo de todas las categorías, es decir, en Perú predominan los establecimiento de primer nivel con una baja calidad.

A nivel de los tipos de suelo, se concluye que se debe buscar una mejor manera de identificar qué tan vulnerable es una vivienda construida en ciertos tipos de suelo.

En el caso de la encuesta de la ENAHO, se observó que la proporción de viviendas construidas con materiales de calidad resistente ante sismos son menores en un treinta por ciento frente a las edificaciones que son vulnerables ante actividades sísmicas.

Por último, se recomienda una mejora en el tratamiento de los datos usados para el entrenamiento del modelo de Clustering. En ese sentido, se requiere un análisis a nivel de variables, con enfoque a las categóricas, para asignar manualmente el peso para cada clase que tenga.

Como trabajos futuros, se propone elaborar rutas de evacuación, estimar la expansión de los sismos y conocer cómo afecta a zonas cercanas. Asimismo, se propone una aplicación del estudio a otros desastres naturales como huaicos o inundaciones.

References

[1] Instituto Geofísico del Perú. (s.f.). Descargar Datos. https://ultimosismo.igp.gob.pe/descargar-datos-sismicos

[2] Morales, A. (s.f.). Cómo realizar geocodificación con GeoPy.

[3] Relación de establecimientos de salud de primer nivel de atención en el Perú. (2023). www.gob.pe. https://www.gob.pe/institucion/minsa/informes-publicaciones/391864-establecimientos-de-salud-de-primer-nivel-de-atencion-en-el-peru

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] K.G. Al-Hashedi, P. Magalingam, Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019, Comput. Sci. Rev. 40 2021.

[9] J.R. Saura, D. Palacios-Marqu´es, D. Ribeiro-Soriano, Using data mining techniques to explore security issues in smart living environments in Twitter, Comput. Commun. 179 2021.

[10] P. Butka, P. Bednár, y J. Ivančáková. Methodologies for Knowledge Discovery Processes in Context of AstroGeoInformatics. In Knowledge Discovery in Big Data from Astronomy and Earth Observation (pp. 1-20). Elsevier, 2020.

[11] ”Escenario de riesgo por sismo de gran magnitud seguido de tsunami frente a la costa central del Perú,” Informe Técnico, Centro Nacional de Estimación, Prevención y Reducción del Riesgo de Desastres, San Isidro - Lima – Perú, 2020.

[12] ”Estrategia de Reducción del Riesgo Sísmico de Edificaciones Escolares Públicas del Perú,” Informe Técnico, Banco Mundial, Washington, DC, 20433 EE. UU, 2017.