**Memoria Análisis Exploratorio de Datos**

**Título del trabajo:** FACTORES GEOESTRUCTURALES QUE PODRÍAN PREVENIR EL COLAPSO DE EDIFICIOS: APLICACIÓN AL SISMO DE GORKHA, NEPAL (2015)

**Principales hipótesis planteadas en el trabajo:**

* El tipo de **adosamiento** influye en la **estabilidad** del edificio: cuanto mayor adosamiento posea un edificio, este será **más resistente** ante las sacudidas de un sismo
* El **número de pisos** de un edificio es clave para conocer si se derrumbará durante un terremoto: **cuanto más alto** sea el edificio, **más susceptible será al colapso**

**Análisis de la primera hipótesis planteada**

Tanto en esta como en la segunda hipótesis, se estudiarán los efectos del terremoto de Gorkha en 4 distritos nepalíes: Gorkha, Dolakha, Makwanpur y Rasuwa (Figura 1).

**Figura 1:** Mapa de la zona central de Nepal, donde se han marcado los 4 distritos incluidos en los análisis estadísticos. Se han marcado los dos epicentros (terremoto principal-izquierda y réplica principal-derecha) y la capital (Katmandú)

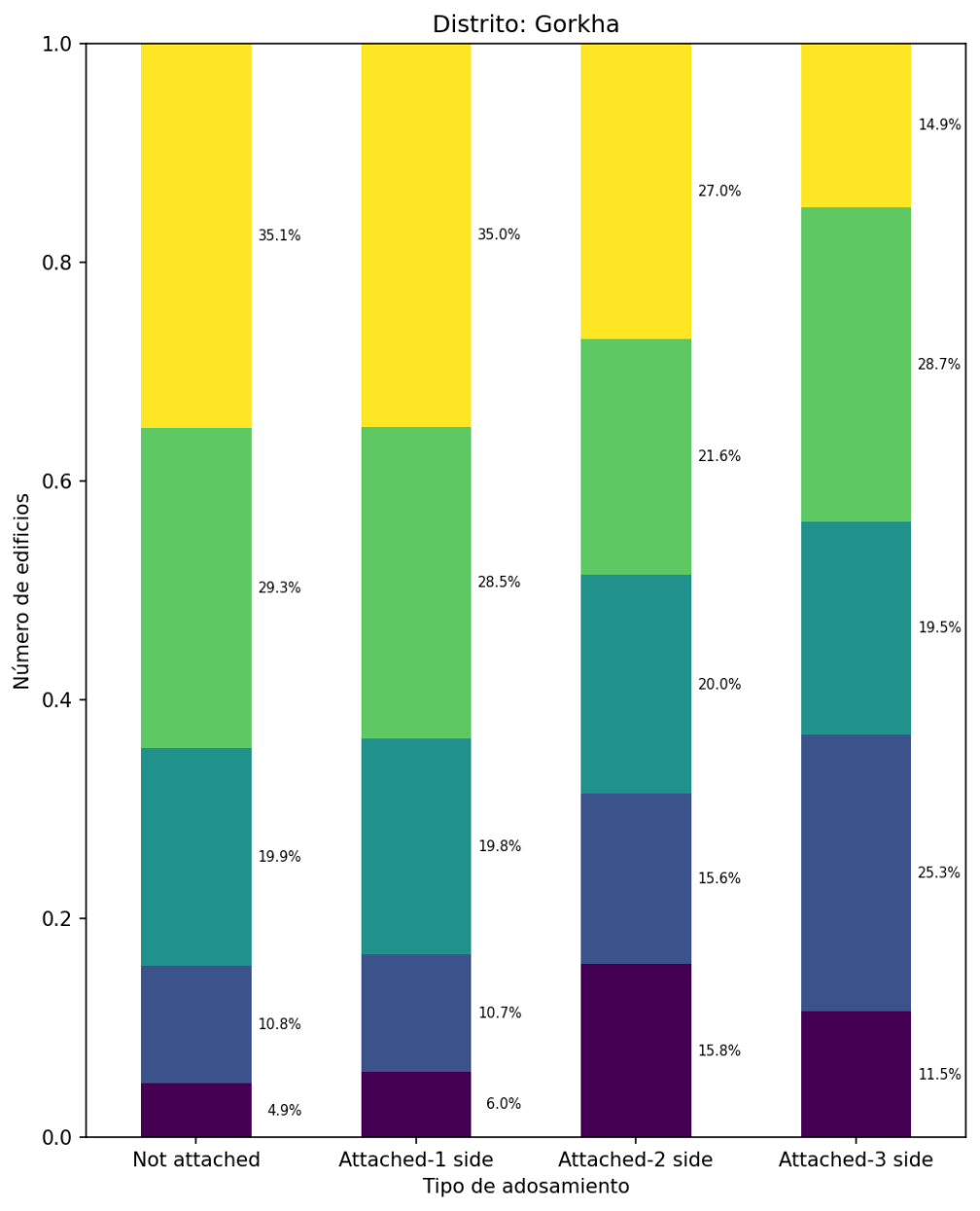
**Rasuwa**

**Dolakha**

**Gorkha**

**Makwanpur**

En este primer análisis estadístico busca dar validez a la hipótesis de que el adosamiento que tiene un edificio indica su resistencia ante las sacudidas de un terremoto. En el dataset del terremoto de Gorkha tenenos una gran cantidad de edificios que se subdividen en “Sin adosar” (Not attached), “Adosado a 1 lado” (Attached 1-side), “Adosado a 2 lados” (Attached 2-side) y “Adosado a 3 lados” (Attached 3-side).

****Una vez cargado el dataset en nuestro archivo .ipynb, con ayuda de un programa creamos barplots normalizados que reflejan la proporción de edificios dañados en función de su adosamiento (Figura 2).

**Figura 2:** Barplots normalizados de la proporción de edificios en el distrito de Gorkha que sufrieron daños durante el terremoto en función de su tipo de adosamiento. Esta proporción está subdividida en 5 grados de destrucción, siendo 1 daños leves a la estructura y grado 5 destrucción total del edificio.

En un primer vistazo, se puede observar como la proporción de edificios derrumbados cae mientras más adosado esté el edificio. Esta apreciación podría ser suficiente para validar nuestra hipótesis, sin embargo, la cantidad de edificios con un adosamiento en 2 o 3 lados es sustancialmente menor que la cantidad de edificios no adosados. Por lo tanto, los resultados obtenidos en altos niveles de adosamiento pudieron haberse dado por azar y puede que no sean representativos. Por ello, haremos un contraste de hipótesis para poder refutar la hipótesis nula. Nuestras dos hipótesis para este contraste de hipótesis son:

**H₀ (nula): El adosamiento y el grado de daño NO están relacionados entre categorías.**

**H1 (alternativa): Existe una asociación significativa entre el grado de daño y el adosamiento entre categorías.**

Al ser las dos variables estudiadas en este contraste de hipótesis (Grado de destrucción y Tipo de adosamiento) dos variables categóricas, el tipo de test que realizaremos será el Chi-cuadrado. Se compararán los valores de los edificios sin adosamiento con los que lo poseen. Asimismo, obtendremos el p-valor y, en caso de ser menor a un umbral preestablecido (0,05), podremos rechazar la hipótesis nula.

Además del valor Chi-cuadrado, al tener pocos datos en el caso de los edificios con adosamientos de 2 y 3 lados, se aplicará el test exacto de Fisher, muy útil para dar robustez a los resultados (este test de Fisher es ideal cuando hay pocos datos por comparar).

Finalmente, el Odds ratio en este caso representa el ratio entre edificios colapsados y no colapsados (edificios con Grade 5 VS edificios en los otros 4 Grades).

**Interpretación del Odds ratio:**

* OR > 1 🡪 el grupo del numerador (aquí Not attached) tiene mayor riesgo.
* OR < 1 🡪 el grupo del numerador tiene menor riesgo.
* OR = 1 🡪 ambos tienen el mismo riesgo.

**Figura 3:** Tabla con los valores de Chi-cuadrado, el p-valor respectivo, p-valor asociado al test exacto de Fisher y el odds ratio de los distritos de Gorkha y Dolakha.

Para los distritos de Gorkha y Dolakha (Figura 3), los p-valores para cada comparación son excepcionalmente bajos especialmente entre los edificios no adosados y los edificios adosados a dos lados. Esto nos indica que los valores de grado de destrucción de los edificios adosados a 2 lados están altamente relacionados con los valores de los edificios sin adosar. En el caso de las otras dos comparaciones los p-valores son también menores que 0,05, con lo que todos los adosamientos no se dieron al azar y se relacionan significativamente con los valores de los edificios sin adosar.

Como la cantidad de edificios con 1 lado adosado no es tan inferior comparada con los edificios sin adosar, no se realizará en este caso el test de Fisher, pues no son cantidades muy pequeñas por comparar.

En general, los p-valores por Fisher también son menores a 0,05 salvo en la relación Not attached/Attached 3-side, de modo que en este caso no se puede rechazar la hipótesis nula. Esto invalida el Odds ratio de 2,88 calculado como suplemento al p-valor de Fisher.

Los valores de Odds ratio son todos mayores a 1, de modo que los edificios sin adosar tienen mayor susceptibilidad al colapso que los demás tipos de adosamientos. (Ejemplo: en el distrito de Gorkha, los edificios sin adosar tienen 3 veces más probabilidad de colapsar que los edificios con 3 adosamientos 🡪 Odds ratio de 3,08).

Realizamos este contraste de hipótesis en los otros dos distritos y obtenemos lo siguiente:

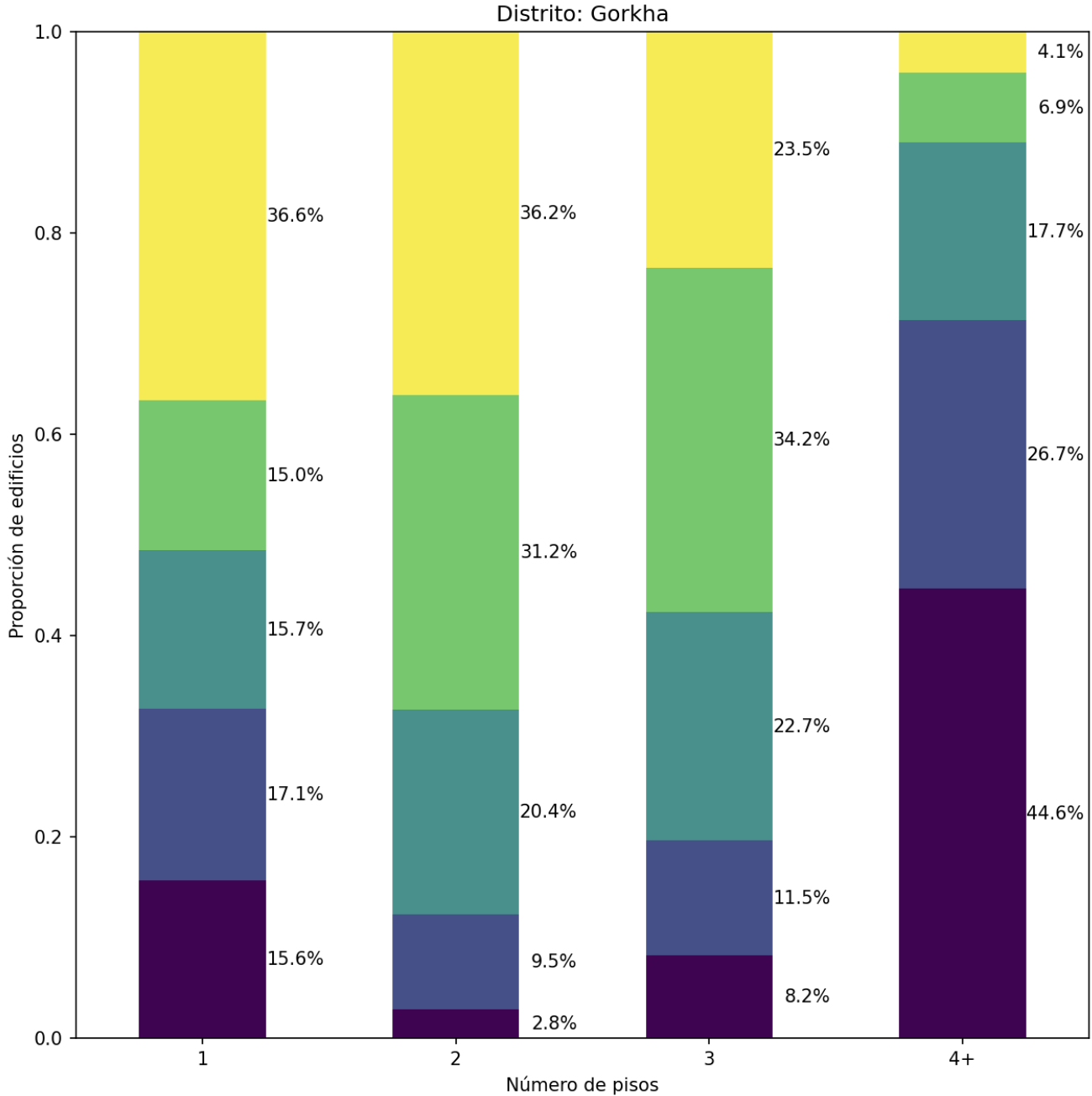
**Figura 4:** Tabla con los valores de Chi-cuadrado, el p-valor respectivo, p-valor asociado al test exacto de Fisher y el odds ratio de los distritos de Makwanpur y Rasuwa.

En los dos siguientes distritos se obtienen también p-valores muy inferiores a 0,05 en el caso de la relación Not attached y Attached 1 y 2-side. Para los edificios con 3 lados adosados, el valor se acerca más al umbral pero no logra rebasarlo, de modo que se puede rechazar la hipótesis nula en todas las relaciones de estos dos distritos (Figura 4).

Los Odds ratio para estos dos distritos también reflejan que los edificios sin adosar tienen mayor probabilidad de derrumbarse que los que poseen adosamiento. (hasta una relación de 5.5 con respecto a Not attached en el distrito de Rasuwa).

Con todos estos datos, podemos concluir que la primera hipótesis quesa validada, ya que hemos podido observar que a mayor adosamiento, mayor es la probabilidad de derrumbarse del edificio (Figura 4), además de que la proporción de edificios colapsados decrece mientras mayor sea la cantidad de adosamientos que posea el edificio (Figura 2).

**Análisis de la segunda hipótesis planteada**

Para este segundo análisis estadístico, se busca validar la hipótesis que busca relacionar una creciente tendencia al colapso de los edificios cuanto mayor sea el número de plantas que lo caracterice. Para ello, en primer lugar, con el fin de dar una mejor visualización de los datos, se agruparán los edificios en 1, 2, 3 y edificios de 4 o más plantas (Figura 5).

**Figura 5:** Barplots normalizados de la proporción de edificios en el distrito de Gorkha que sufrieron daños durante el terremoto en función del número de plantas del mismo. Esta proporción está subdividida en 5 grados de destrucción, siendo 1 daños leves a la estructura y grado 5 destrucción total del edificio.

A primera vista, se constata que, contrariamente a lo planteado en la hipótesis, a partir de 4 o más pisos, la susceptibilidad al colapso desciende incluso por debajo de los edificios con 1 único piso.

Para poder corroborar esta observación, haremos como en la primera hipótesis. Comprobaremos que la pequeña cantidad de edificios de 4 o más plantas nos han ofrecido datos interpretables o son puramente debidos al azar. Así pues, el contraste de hipótesis tendrá las siguientes hipótesis:

**H₀ (nula): El número de pisos y el grado de daño NO están relacionados entre categorías.**

**H1 (alternativa): Existe una asociación significativa entre el grado de daño y el adosamiento entre categorías.**

En el análisis de esta segunda hipótesis también compararemos dos variables categóricas, de modo que usaremos de nuevo el test de Chi-cuadrado.

Además de calcular el p-valor para cada tabla de contingencia (una tabla por cada distrito), se creará un heatmap con los valores residuales estandarizados (Figura 6).

Cada celda del heatmpa representa un residuo estandarizado (Z) entre el número observado (de edificios dañados) y el número esperado.

Si el valor de Z es mayor a 2 o menor que -2, el valor de la celda se tiene en cuenta como significativo. La intensidad del color en el heatmao indica la gran diferencia entre valores observados y esperados. Si es color rojo intenso, hay más edificios con ese grado de daño que los esperados y viceversa si el color de la celda es azul intenso.

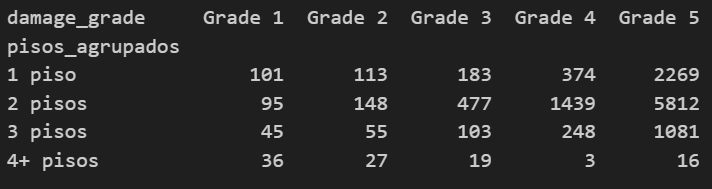
En el caso del distrito de Rasuwa, se puede observar una tendencia decreciente de edificios dañados para los edificios de 4 o más pisos. Hay muchos más edificios con muy poco daño (Grade 1) y bastante menos cantidad de edificios que colapsaron (Grade 5). Esto nos indica que los edificios de 4 o más plantas en el distrito de Rasuwa son más resistentes al derrumbamiento de lo esperado.

Asimismo, en este mismo distrito, hay menos edificios de 2 plantas con daños leves de lo esperado (Grade 1). El valor de Z va aumentando poco a poco al ir incrementándose el número de pisos, de manera que el valor de 3,82 para un grado de destrucción de 4 nos indica que hay más edificios de dos plantas que sufrieron daños muy severos en Rasuwa.

Estos resultados afianzan las observaciones previas sobre la tendencia de los edificios de muchas plantas a ser más resistentes contra sacudidas sísmicas si lo comparamos con el comportamiento de los edificios de 2 o 3 pisos.

Para esta tabla de contingencia del distrito de Rasuwa (Figura 7), se ha obtenido un p-valor asociado al test de Chi-cuadrado de 3,95 e-197, siendo un valor tan minúsculo que podría considerarse cercano a 0. Con este p-valor, se puede rechazar cómodamente la hipótesis nula, con lo que en este distrito se comprueba que el número de pisos y el grado de destrucción están fuertemente relacionados, cuyos datos no parecen haberse dado debido al azar.

**Figura 6:** Heatmap de los residuos estandarizados (Z) del test de Chi-cuadrado en el distrito de Rasuwa. A mayor valor de Z, más significativo será la proporción de edificios dañados observados con respecto a los esperados.



**Figura 7:** Tabla de contingencia con el número de edificios del distrito de Rasuwa clasificados por número de pisos y grado de destrucción sufrido.

Este proceso se repetirá en los 4 distritos, obteniendo valores muiy diversos. En todos los dtistritos el p-valor es menor a 0,05, con lo cual se puede concluir que en los distritos estudiados existe una relación significativa entre el número de pisos y el grado de daño que sufrieron durante el terrmoto de Gorkha.

**Observaciones y conclusiones del EDA**

1. **PRIMERA HIPÓTESIS**

El adosamiento de los edificios está relacionado significativamente con el grado de destrucción que sufrieron tras el terremoto de Gorkha.

Los edificios sin adosar poseen una susceptibilidad al derrumbamiento entre un 1.4 y un 3.85 mayor con respecto a los edificios con 2 lados adosados.

Los edificios sin adosar poseen una susceptibilidad al derrumbamiento entre un 2.79 y un 5.5 mayor con respecto a los edificios con 3 lados adosados.

**Se CONFIRMA nuestra hipótesis ⟶ A mayor adosamiento ⟶ mayor resistencia**

1. **SEGUNDA HIPÓTESIS**

El número de pisos está relacionado significativamente con el grado de destrucción que sufrieron tras el terremoto de Gorkha.

* **En el distrito de Rasuwa:**

**⟶ mayor resistencia al colapso de los edificios de 4 o más plantas**

**NO SE CONFIRMA nuestra hipótesis ⟶ A partir de 4 plantas es más resistente que 2 o 3 pisos**

**Dificultades encontradas en este trabajo**

La principal dificultad en este trabajo ha sido la obtención de los datos. Existen pocas fuentes de datos referentes al terremoto de Nepal, además de que el acceso a información sobre los distritos y el id asociado a estos fue muy complicada ya que la mayoría de las webs gubernamentales nepalís están caídas o no poseen prácticamente nada de información. Gracias a una base de datos de Kaggle, pude encontrar una base de datos con una gran cantidad de edificios disponibles (más de 760 000) y una amplia varidad de características intrínsecas a cada edificio con los que poder realizar un análisis exhaustivo de estos datos.

Asimismo, el manejo de Folium ha sido en ciertos puntos muy engorroso por la novedad de esta herramienta.

Finalmente, al aplicación del heatmap de residuos estandarizados fue un poco compleja al principio por la aplicación al trabajo y la novedad en comparación con los heatmpas vistos en clase.

**Documentación utilizada para realizar el EDA:** <https://www.kaggle.com/datasets/sanskarnegi/nepal-earthquake-2015>