

Predicción de la percepción de sismos por la población a partir de datos básicos de terremotos

Trabajo final de la asignatura
“Taller de tesis”



**Maestría en Explotación de Datos y
Descubrimiento del Conocimiento**



Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

Autor:

Víctor A. Bettachini

Fecha:

18 de junio de 2024

Resumen

Una decena de terremotos se producen en el territorio nacional en forma diaria sin que sean detectados más que por instrumentos. Son muy pocos los casos en que las ondas sísmicas que estos producen son percibidos por la población. Este trabajo busca cuantificar la capacidad de predecir esta eventualidad por dos herramientas de la ciencia de datos usando solo datos básicos de terremotos.

Índice general

1. Introducción	3
1.1. Contexto y motivación científica	3
1.2. Objetivos del trabajo / Pregunta	4
1.3. Estructura del documento	4
2. Marco teórico	5
2.1. Relevamiento de trabajos previos y relevantes	5
2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo	6
3. Metodología	7
3.1. Presentación y descripción de los datos utilizados	7
3.2. Limpieza de los datos	8
3.2.1. Modificaciones de los datos (ingeniería de características)	9
3.3. Análisis exploratorio de datos (AED)	10
3.4. Descripción de las técnicas de análisis y, si corresponde, de mo- delado	13
3.5. Descripción de la selección de características (si corresponde) . .	13
3.6. Preprocesamiento	15
3.7. Descripción de las métricas de evaluación de los modelos (si co- rresponde)	16
3.8. Descripción de los métodos estadísticos utilizados (si corresponde)	16
3.8.1. Regresión logística para la predicción binaria	16
3.8.2. XGBoost para una predicción binaria	16
4. Resultados y discusión	17
4.1. Presentación y análisis de resultados obtenidos	17
4.1.1. Regresión logística	17
4.2. Discusión de los resultados y su relevancia	17
4.3. Limitaciones y posibles mejoras	17
5. Conclusión	18
5.1. Resumen de los hallazgos principales	18
5.2. Conclusiones generales y su relación con los objetivos del trabajo	18
5.3. Aplicaciones y relevancia de los resultados	18
Bibliografía	19

Anexos (opcionales)	21
5.4. Código fuente utilizado en el análisis (link a un repositorio) . . .	21
5.5. Tablas y gráficos adicionales	21
5.6. Otros materiales relevantes	21

Capítulo 1

Introducción

1.1. Contexto y motivación científica

Un rápido desprendimiento entre dos fascetas enfrentadas de sendas placas tectónicas que traban mutuamente su desplazamiento relativo produce una rápida liberación de energía que se denomina terremoto. Esto sucede a cierta profundidad en la corteza terrestre en el punto denominado hipocentro a partir del cual parte de esta energía se encauza como ondas elásticas. El estudio de estas ondas es el área llamada sismología y de ahí el término sismo para un evento particular detectado, pero que debiera aclararse si se produjo por un terremoto u otra fuente de ondas [1, sección 4.1.1]. Sea que las ondas sean de compresión longitudinal de la corteza, las tipo P, o las del tipo S transversales y más lentas, ambas arriban con mayor intensidad al punto de la superficial terrestre que se encuentra directamente sobre el hipocentro, que se denomina epicentro [1, sección 4.1.2] como ilustra la figura 1.1. Cuanto más próxima es una locación en la superficie a un epicentro la amplitud de las ondas sísmicas es mayor. Tanto esta amplitud como el período de oscilación son mucho mayores que el de otros desplazamientos de la corteza respecto a otros desplazamientos como los de las mareas solares y lunares de la corteza [1, sección 4.1.4]. Como resultado estructuras artificiales pueden agitarse poniendo en riesgo su estabilidad estructural y haciendo caer elementos que no estaban fijados a esta o perdieron tal adhesión a causa de la agitación misma. Como consecuencia los sismos más fuertes pueden generar graves daños, poniendo en riesgo la integridad física y la seguridad de las personas al generar daños en las viviendas y edificios, derrumbes de puentes, rompimiento de vidrios, entre otros [2].

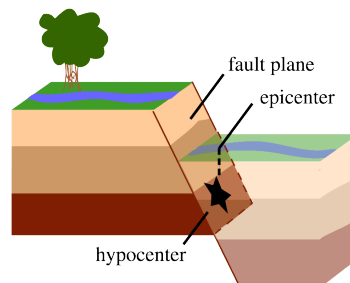


Figura 1.1: Epicentro e hipocentro de un sismo.

Pero de hecho la mayor parte de los sismos son de muy pequeña magnitud y no generan daños materiales. Sin embargo el que sean o no detectados por la población es un factor relevante en su percepción de confianza vis-à-vis de

los organismos de monitoreo y prevención de riesgos. Sí o cuando informar a la población de la ocurrencia de un sismo es una decisión de política pública que debiera apuntar a no alertar innecesariamente sobre sismos menores imperceptibles [3]. El caso inverso es también problemático con casos en que ante una carencia de una comunicación oficial de la poca importancia de un evento sísmico llevó a la autoevacuación por parte de la población que lo percibió [4].

1.2. Objetivos del trabajo / Pregunta

Contar con una estimación rápida a partir de los datos sísmicos registrados por instrumental de si un dado evento será percibido por la población o no permitiría a las autoridades tomar decisiones informadas sobre la comunicación a la población. Este trabajo busca determinar el grado de certeza con que ciertos métodos de ciencias de datos pueden predecir si la población percibirá actividad sísmica producto de terremotos a partir de unos pocos datos básicos sobre los mismos publicados por el Instituto Nacional de Prevención Sísmica (INPRES) que es el organismo público de Argentina que realiza estudios e investigaciones básicas y aplicadas de sismología [5].

1.3. Estructura del documento

Se estructuró en los siguientes capítulos con temáticas diferenciadas

Introducción : se presentará el contexto y la motivación científica, los objetivos del trabajo y la estructura del documento.

Marco teórico : se revisarán los trabajos previos y relevantes, se presentarán los conceptos y técnicas de ciencia de datos utilizados en el trabajo.

Metodología : se describirán los datos utilizados, el preprocesamiento y limpieza de los mismos, el análisis exploratorio de los datos, las técnicas de análisis y modelado utilizadas, la selección de características, las métricas de evaluación de los modelos y los métodos estadísticos utilizados.

Resultados y discusión : se presentarán y analizarán los resultados obtenidos, se discutirán los resultados y su relevancia, se identificarán las limitaciones y posibles mejoras.

Conclusión : se resumirán los hallazgos principales, se presentarán las conclusiones generales y su relación con los objetivos del trabajo, se discutirán las aplicaciones y relevancia de los resultados.

Capítulo 2

Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

En relación con la temática de terremotos los trabajos publicados que hacen uso de herramientas de aprendizaje automático en su mayoría persiguen el esquivo objetivo de predecir la ocurrencia de sismos de gran intensidad, y en segundo lugar para la mejora de la detección de estos por parte de instrumental, pero la temática de la percepción de los mismos por humanos no parece ser una temática corriente [6].

Un trabajo interesante no porque sea un antecedente del enfoque o temática de este trabajo pero sí por tocar al aspecto social de la percepción de sismos hace uso del método conocido como SHAP (contracción de SHapley Additive exPlanations, explicaciones aditivas de Shapley) que determina la contribución de cada característica en un conjunto de datos a una determinada predicción [7]. Los datos son características sociológicas de los individuos y la predicción es sobre la percepción personal del riesgo sísmológico. Los autores muestran un resultado estadístico que avala el que se incrementa la percepción del riesgo de esta fuente tras haber percibido un sismo [8].

Un estudio de la problemática de la percepción por la población de un terremoto tal vez deba separarse en dos problemas a resolver secuencialmente. En primer lugar determinar el espectro de frecuencias de oscilación del suelo en función de características físicas del terremoto y la corteza. Y en segundo lugar la respuesta de las personas a tal espectro de oscilaciones que depende además de características de las viviendas más crucialmente de cuestiones fisiológicas.

Atendiendo a la primer cuestión la modelización física ha arribado a un conjunto de ecuaciones para predecir el espectro de frecuencias de oscilación del suelo en función de la distancia al hipocentro y las características de la corteza en el camino de propagación de las ondas desde el terremoto [9]. Pero como estas ecuaciones dependen de una multitud de factores que deben ajustarse a cada locación en muchos casos en forma empírica se han ensayado alternativas basadas en el aprendizaje automático en años recientes, en particular utilizando arquitecturas de redes neurales recurrentes [10].

En cuanto a la segunda cuestión, la respuesta fisiológica al espectro de oscilaciones del suelo, el foco también ha sido el consante en el campo de la sismología,

el de la predicción de terremotos de gran importancia. El campo se ha enfocado primordialmente a la capacidad sensora no de humanos sino de animales tradicionalmente considerados más sensibles y fuentes de señal de alarma ante sismos [11].

Pero en la superficial búsqueda superficial de literatura realizada para este trabajo no encontré antecedentes que cubrieran todo el arco desde características que describen el terremoto hasta la determinación excluyente de si los correspondientes sismos son o no percibidos por la población.

2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo

Siendo el objetivo de cualquier modelo utilizado una clasificación entre dos clases, la percepción de un sismo o no, se busca determinar que tan fuerte es el vínculo de cada una de las variables de los sismos de las que se dispone datos y la percepción de los mismos por parte de la población. Para esto dos técnicas se planea utilizar dos distintas técnicas de la ciencia de datos: la regresión logística y la clasificación de árboles de decisión.

Un modelo de regresión logística permite una clasificación binaria de los datos en función de una variable dependiente y un conjunto de variables independientes al tiempo de dar un peso a cada una de estas últimas lo que permitirá transmitir cuales son más relevantes en la clasificación. Determinar la atribución de la percepción de un sismo a las variables de los mismos se considera de interés para los objetivo de este trabajo. Ajustando el grado de regularización tipo Lasso (L1) se buscará reducir el número de variables independientes en el modelo verificando en que grado su omisión afecta a la predicción.

Por otra parte se buscará utilizar una herramienta más reciente, las máquinas de potenciación de gradiente, más conocidos por su nombre en inglés Gradient Boosting Machines (GBM). Las distintas implementaciones de estos algoritmos, como XGBoost, LightGBM o CatBoost son capaces de producir un único modelo con fuerte poder predictivo a partir de la síntesis de resultados de modelos de predicción débiles, típicamente árboles de decisión.

Puesto que los GBM carecen de un mecanismo para evidenciar la importancia de las variables en la clasificación como la que evidencian los pesos de la regresión logística, se planea utilizar los valores del método SHAP comentado en la sección anterior para tal fin. Para el lenguaje R está disponible la biblioteca *shapr* para tal fin [12].

El disponer de una herramienta de explicación de los modelos de predeción aplicable al resultado de los dos a ensayar, el de regresión logística y el XGBoost, permitirá comparar la relevancia de las variables que asignará cada uno para la percepción de los sismos por parte de la población.

Capítulo 3

Metodología

3.1. Presentación y descripción de los datos utilizados

En el marco de los “Proyectos de Asistencia Estadística del Instituto de Cálculo (IC)” de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad de Buenos Aires (UBA) se publicaron conjuntos de datos en un repositorio curado con el objeto de ser aplicados a la enseñanza de la estadística y la ciencia de datos por Daniela Parada, investigadora del IC [13]. De estos conjuntos el utilizado en este trabajo es el que se publica en el apartado “Visualización” que corresponden a datos de sismos de Argentina de la última década [14]. En este repositorio alojado por la firma GitHub, se provee un front-end html que da un contexto, hace una exploración inicial, un análisis para una provincia en particular, muestra una estimación de probabilidad y provee otra información sobre lo datos.

Los datos corresponden a detecciones por parte de estaciones de monitoreo sísmico en la República Argentina recopilados y publicados por el INPRES en su sitio web [15]. En el sitio de publicación de los datos se indica que el conjunto de datos comprende las fechas desde el 7 de enero de 2012 hasta el 18 de mayo de 2022 y fue realizado con datos *scrappeados* del buscador de sismos del INPRES por Gustavo Juantorena [14, sección 4.1].

Allí mismo se describe que el conjunto de datos reducido y curado denominado “sismos”, el que se utilizó en este trabajo, es accesible a través de la importación de la biblioteca `datosIC` en lenguaje R [14, sección 5.1.1]. Este mismo conjunto reducido puede descargarse en formato de texto separado por comas (CSV) apuntando a su URL en el repositorio GitHub¹.

Las variables reportadas para cada sismo son:

- *Fecha*: en el formato `verb‘aaaa-mm-dd’` de la norma ISO 8601 [16].
- *Hora*: una cadena de caracteres en formato `hh:mm:ss` con una exactitud al segundo.

¹https://github.com/daniellaparada/IC-datasets-docencia/blob/main/fuente/04_visualizacion/sismos-arg.csv

- *Latitud, Longitud*: un número con una exactitud de un decimal con grados como unidad.
- *Provincia*: cadena de caracteres del nombre de la provincia donde se produjo el sismo (no donde se ubicó quién potencialmente lo percibiera) según se afirma en el sitio de publicación [14, pág. 5.1.1].
- *Magnitud*: un número con la escala Richter como unidad una función de la amplitud de las ondas sísmicas [1, sección 4.2.3]
- *Profundidad*: un número entero con kilómetros como unidad que indico que tan bajo la superficie se ubicó el epicentro.
- *Percibido*: variable booleana de si hubo reportes de percepción del fenómeno por parte de la población,

Esta última variable es la que se busca predecir en este trabajo en función de las demás.

3.2. Limpieza de los datos

Puesto que el conjunto de datos es curado por un equipo de investigación de la UBA, se asume que los mismos son confiables y que no se requiere de un proceso de limpieza de los mismos. De todas formas se realizaron las verificaciones usuales cada vez que se utilizan datos tabulares en un estudio de estadístico y/o de ciencia de datos.

Carga de los datos Tras descargar el archivo de datos en fomato CSV se le importó en una estructura de datos *data.table* de en un entorno de trabajo en lenguaje R. Esta estructura de datos permite una consulta de los datos análoga a la del lenguaje SQL de bases de datos relacionales lo que le hace una herramienta versatil para el análisis de datos tabulares [17].

Inspección de tipos de datos Ejecutar la función `colnames` con la *data.table* denominada `sismos_arg` como argumento permitió verificar que contuviera las columnas con los nombres anunciados en el sitio que publica los datos en su sección [14, Exploración inicial]. Las transformaciones o ingeniería de características que se detallan luego en esta sección se realizaron en función de los tipos de datos de cada columnas constatados con la función `str`.

Valores faltantes o duplicados La presencia de valores faltantes indicados con el símbolo NA se descartó cuando la ejecución `sum(is.na(sismos_arg))` arrojó un cero como resultado. Por el contrario unos 23 registros duplicadas mostró ejecutar `sismos_arg[duplicated(sismos_arg, fromLast = TRUE)]` sobre un total de 55817 registros. Se hizo una copia de la tabla sin registros duplicados ejecutando `sismos_arg[!duplicated(sismos_arg)]` en una nueva tabla con nombre más corto, `sismos`.

Datos atípicos La detección y potenciales acciones sobre datos atípicos se tratan una vez iniciado un análisis exploratorio de datos temática de la sección 3.3.

3.2.1. Modificaciones de los datos (ingeniería de características)

Por otra parte el preprocesamiento comprende la generación de nuevas variables a partir de las existentes que se consideren relevantes para el análisis, lo que recibe el nombre de ingeniería de características (feature engineering). Se generaron dos nuevas columnas a partir de las existentes en el conjunto de datos original, una en función al formato de los datos en la columna `Hora` y otra en función de la escala física utilizada en la columna `Magnitud`.

Decimalización de la hora del sismo La columna “Hora” está codificada como una cadena de caracteres. Para poder utilizarla en un análisis de regresión se la convirtió a un número entero de segundos transcurridos desde la medianoche del día en que se produjo el sismo. Para ello se escribió una función `convert_to_seconds` centrada en la función `strptime` del paquete base de R. Se verificó que la columna generada con aplicación a través de `sismos[, Hora_decimal := sapply(Hora, convert_to_seconds)]` estuvieron en el rango de 0 a 86 400 s, es decir, el de un día completo.

Recorte por profundidad Puesto que hay sismos que tienen su origen en desplazamientos de tierra, explosiones artificiales para la minera, fracturado hidráulico para la extracción de hidrocarburos que no son terremotos, se omiten los datos informados con una profundidad de 0 km con simple un filtrado `sismos <- sismos[Profundidad > 0]`. Esta es una práctica usual en el análisis de datos orientados a sismos originados en terremotos [6]. Siendo que la variable se informa como enteros de kilómetros, estos representarían los hipocentros hasta una profundidad de 500 m, compatibles con estas actividades artificiales. Esto es un número muy pequeño de los sismos en el conjunto de datos como la muestra el histograma según su profundidad que reproduce la figura 3.1.

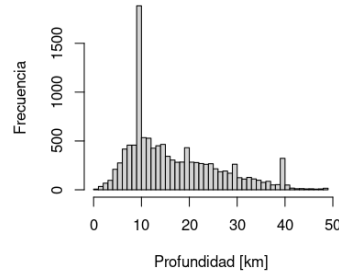


Figura 3.1: Los sismos con profundidad de 0 km son pocos.

Linealización de la magnitud Todas las escalas de magnitud de los sismos son logarítmicas con una forma genérica

$$M = \log_{10} \left(\frac{A}{T} \right) + q(\Delta, h) + a, \quad (3.1)$$

donde M es la magnitud, A es la amplitud de las ondas, T es el período de oscilación, q es una función de ajuste para el ángulo desde la vertical desde el sismógrafo al Δ y su profundidad h , y a es una constante de ajuste [1, ecuación 4.13]. Son las ondas S y P (ver sección 1.1) tienen distintos coeficientes de ajuste en la ecuación 3.1 y por ello se utilizan distintas escalas de magnitud. La más generalmente usada hoy día es la de Gutenberg-Richter publicada en 1956 para la magnitud de este tipo de ondas, m_b ,

$$m_b = \log_{10} \left(\frac{A}{T} \right)_{\text{máx}} + q(\Delta, h), \quad (3.2)$$

sin la constante de ajuste a y con valores de $q(\Delta, h)$ en un rango ≈ 6 a 8 para ondas P en $\Delta = 10$ a 110° [1, ecuación 4.18]. Entonces la escala de Gutenberg-Richter de un terremoto es el logaritmo en base 10 de la mayor razón de la amplitud y período de las ondas registradas por los sismógrafos incluyendo ajustes para compensar la variación en la distancia entre el sismógrafo y el hipocentro [18].

Puesto que la magnitud busca dar cuenta de la energía liberada en el terremoto pero este trabajo busca predecir la percepción de los sismos, es de interés obtener un valor que sea lineal con la amplitud registrada por los sismógrafos en superficie, A , que también es la que afecta a las personas. Para obtener un valor lineal con esta A a partir de los datos de magnitud del INPRES se consideró usar una valor medio del rango comentado en el párrafo anterior, $\bar{q}(\Delta, h)$, con lo que podría despejarse

$$\left(\frac{A}{T} \right)_{\text{máx}} = 10^{(m_b - \bar{q}(\Delta, h))}. \quad (3.3)$$

Pero si se toma tal promedio se establecería un $\bar{q}(\Delta, h) = 7$, que cuasará que haya valores negativos para el valor a a la izquierda de la ecuación pues el menor m_b en el conjunto de datos es 2,5, y ese sería algo sin validez física (¡amplitudes negativas!). Se utilizará entonces este último como valor para $q(\Delta, h)$.

De querer despejar la amplitud de las ondas, A , debe establecerse que hacer con el período de las ondas, T , dato que no figura en el conjunto de datos. No se encontró otra alternativa que asumir que todos los fenómenos registrados tienen el mismo y en consecuencia asumir $T_{\text{constante}}$. Así realizando despejes a partir de la ecuación 3.2 asumiendo tales condicionantes todo cuanto puede aproximarse a un valor relacionado la amplitud será una lineal con registrada por un sismógrafo, $A_{\text{máx}}$,

$$\left(\frac{A_{\text{máx}}}{T_{\text{constante}}} \right) = 10^{(m_b - 2,5)}. \quad (3.4)$$

Con el comando `sismos[, proxy_amplitud := 10^(Magnitud- 2.5)]` se generó una columna para este valor denominada `proxy_amplitud`.

3.3. Análisis exploratorio de datos (AED)

Un primer vistazo sobre los datos con `summary(sismos)` permitió obtener un resumen de las variables numéricas y categóricas. Saltan a la vista que hay valores extremos en la variable `Magnitud` y un fuerte desbalance en la variable “Percibido”, la de clase de clasificación.

Desbalance de la clase de clasificación Sobre el total de 55794 registros únicos, un $\approx 96,6\%$ solo fueron percibidos por el instrumental y no por la población. Restan tan solo unos 1905 registros, un $\approx 3,4\%$, que si fueron percibidos por la población. Este desbalance llama al uso de técnicas de balanceo de clases en los modelos de clasificación a utilizar.

Distribución de la magnitud La magnitud de terremotos forzosamente presenta una distribución asimétrica por el hecho de que cuanto mayor es la energía liberada, más infrecuente es el fenómeno. Esto la ilustra el número por año de sismos representado en una escala logarítmica en función de la magnitud generado con código provisto junto con los datos [14, sección 4.2.1], que se reproduce en la figura 3.2.

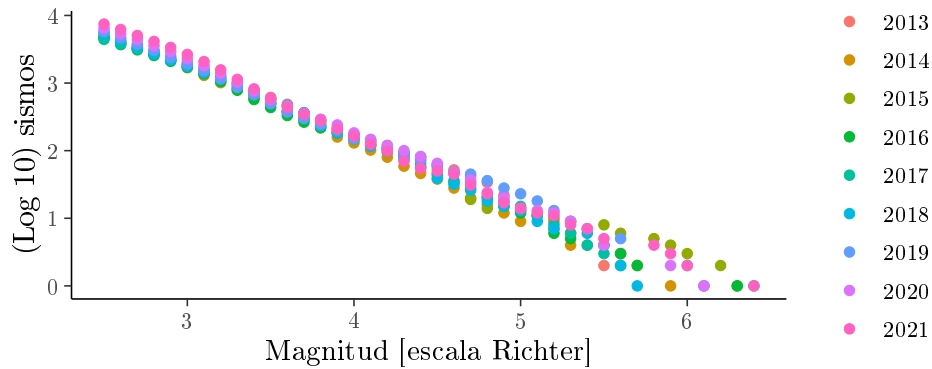


Figura 3.2: Los sismos de mayor magnitud son más infrecuentes.

Horario de los percibidos Puede hipotetizarse que en los horarios de sueño de la mayoría de la población sea menor la proporción de sismos percibidos. Para ver si esto es se segmentaron los fenómenos con la variable `Hora_decimal` en intervalos de una hora. En cada intervalo calculó la proporción entre casos percibidos o no y se graficaron las 24 proporciones generando la figura 3.3.

Contrariando la hipótesis se observa en que si hay una merma en percepciones de sismos esta es muy ligera y se produce entre las 10 y 19 horas. Donde es claro un cambio llamativo es en el incremento durante las tres horas que comienzan a las 23 de cada día. Habría que ver si tal variación es estadísticamente significativa, pero por sobre todo buscar una posible causa sociológica o fisiológica relacionada a que sean las primeras horas de sueño nocturno de la mayor parte de la población.

Distribución geográfica Como es esperable la mayor parte de los hipocentros se ubican en regiones con orografía elevada, como la cordillera de los Andes, producto de la subducción de la placa de Nazca bajo la placa Sudamericana, o las sierras cordobesas producto de procesos mucho más antiguos. Una ubicación de los mismos sobre un mapa físico lo ilustra en la figura 3.4. Esto conlleva a

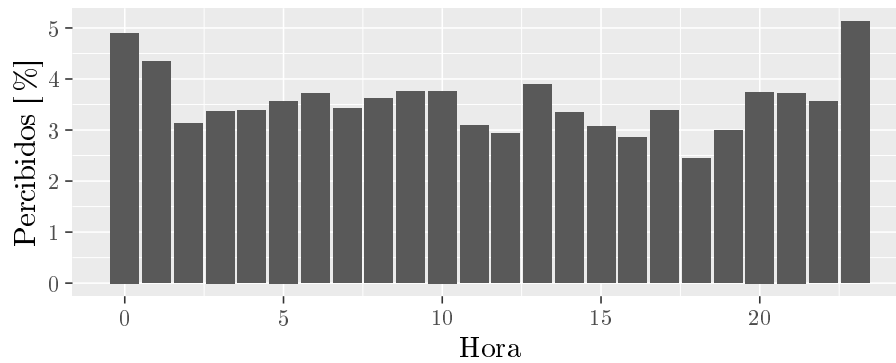


Figura 3.3: Proporción de sismos percibidos por la población en función de la hora del día.

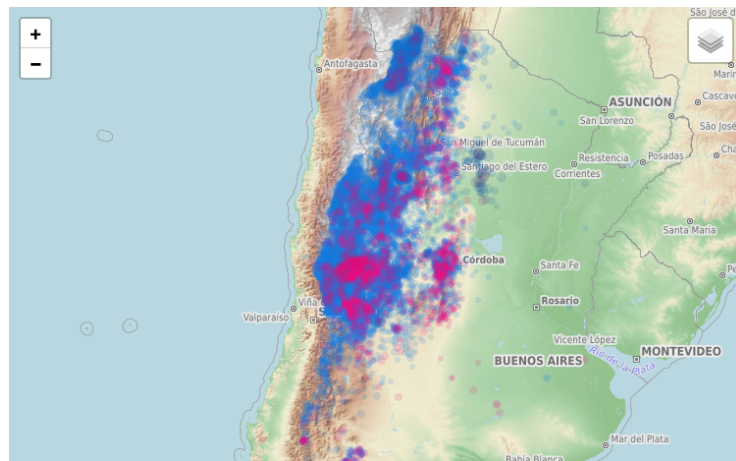


Figura 3.4: Los sismos no percibidos, en azul, predominan pero la mayoría se registran en zonas de montaña o cerros en su mayoría alejados de las mayores urbanizaciones. Reproducido de [14]

que la mayor parte de los sismos se produzcan en zonas de montaña o cerros en su mayoría alejados de las mayores urbanizaciones

La ubicación solo de los percibidos en un mapa político que muestra la figura 3.5 muestra cierta concordancia con la figura anterior al tiempo de dejar a las claras que la percepción de los sismos por la población no es un fenómeno territorialmente homogéneo aún en provincias con abundante actividad sísmica.

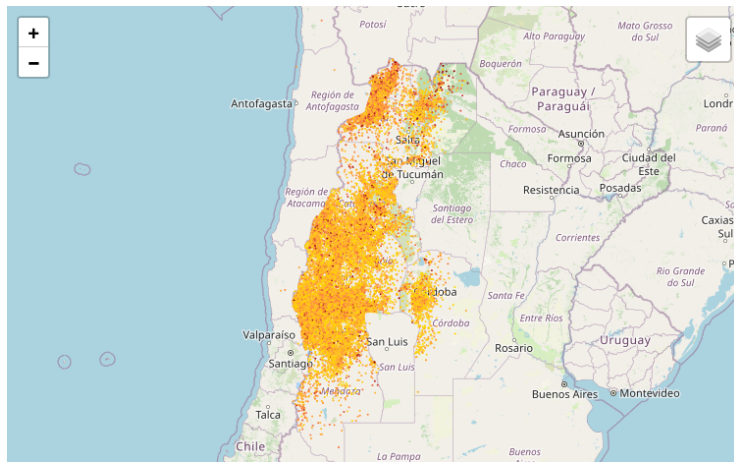


Figura 3.5: Los sismos percibidos tienen una distribución poco homogénea aún en provincias con abundante actividad sísmica.

3.4. Descripción de las técnicas de análisis y, si corresponde, de modelado

3.5. Descripción de la selección de características (si corresponde)

El problema de la distancia Evidentemente la proximidad de un epicentro a una población es un factor relevante en la percepción de un sismo. El INPRES tiene una metodología operativa para registrar la ubicación de los usuarios que reportan haber percibido un sismo a través de su página web [19]. Su buscador de sismos indica los sismos “sentidos” por usuarios con un color en sus resultados de búsqueda como ilustra la figura 3.6.

Obviando latitud y longitud Lamentablemente en el conjunto de datos curados no se dispone de la ubicación de la población que informó haberles percibido lo que imposibilita incorporar la distancia al epicentro como una variable en el modelo sin realizar un nuevo *scrapping* de estos datos en el sitio del INPRES. Por tal razón para este trabajo se obvian las variables de latitud y longitud del epicentro. Para que esto no sea un factor de sesgo en el modelo se decidió filtrar por la variable “Provincia” y trabajar con un subconjunto de los datos que correspondan a una provincia en particular.

Elección de la provincia de San Juan Se busca que esta presente un número elevado de sismos, que tenga una extensión geográfica relativamente limitada y que el reporte de sismos percibidos no tenga sesgos geográficos, es decir una mayor preponderancia en regiones meridionales y/o orientales que en los sentidos opuestos. Como puede apreciarse en la figura 3.5 presentada en la sección 3.3 la provincia de San Juan cumple con estos requisitos. Se eligió entonces la provincia de San Juan por presentar un elevado número de



Sismología	Red Nacional de Estaciones Sismológicas	Red Nacional de Acelerógrafos	Ingeniería Sismorresistente	Laboratorio de Estructuras	INPRES-CIRSOC Reglamento	Publicaciones	Servicios
------------	---	-------------------------------	-----------------------------	----------------------------	--------------------------	---------------	-----------

Buscador de sismos

Resultado de la búsqueda

12sismos encontrados según las características especificadas. (Los sismos listados en color rojo han sido sentidos.)									
Id	Fecha	Hora	Latitud	Longitud	Profund.	Magn.	Intensidad	Provincia	
1	14/06/2024	23:32:49	-31.860	-69.663	119 Km.	2.7		SAN JUAN	mapa
2	14/06/2024	22:12:32	-23.274	-66.653	233 Km.	3.5		JUJUY	mapa
4	14/06/2024	20:15:52	-31.897	-69.960	126 Km.	2.9		SAN JUAN	mapa
5	14/06/2024	18:51:09	-31.547	-69.410	125 Km.	2.8		SAN JUAN	mapa
6	14/06/2024	18:44:59	-28.521	-67.919	112 Km.	2.5		LA RIOJA	mapa
7	14/06/2024	16:12:04	-31.250	-68.664	108 Km.	3.8	II a III -Ciudad de San Juan, San Juan; II a III -Albardón, San Juan; II a III -Caucete, San Juan; II a III -Villa San Martín, San Juan	SAN JUAN	mapa
8	14/06/2024	14:41:57	-31.538	-69.343	102 Km.	2.8		SAN JUAN	mapa
9	14/06/2024	14:02:37	-31.898	-69.009	111 Km.	2.8		SAN JUAN	mapa
10	14/06/2024	12:38:36	-31.818	-69.715	100 Km.	2.5		SAN JUAN	mapa
13	14/06/2024	04:38:59	-31.297	-68.555	108 Km.	2.8		SAN JUAN	mapa
14	14/06/2024	04:15:16	-23.687	-66.592	227 Km.	3.2		JUJUY	mapa
15	14/06/2024	03:30:23	-31.210	-68.433	104 Km.	3.1		SAN JUAN	mapa

1

volver a Búsquedas

Figura 3.6: Resultados de una búsqueda manual de sismos en el sitio del INPRES. El indicado en rojo fue percibido por la población. En la columna *intensidad* se dan datos de ubicación de la población que lo percibió.

casos percibidos o no en el conjunto de datos, por tener una extensión geográfica reducida en comparación con otras así como contar con una alta homogeneidad espacial en el reporte de sismos hace que el factor distancia entre epicentro y población que reporta tenga menor impacto que en otras provincias. Se procedió a filtrar los datos por la provincia de San Juan con el comando `sismos_SJ <- sismos[Provincia == "San Juan"]`. Afortunadamente el subconjunto de datos de esta provincia representa un $\approx 53\%$ de los datos hasta aquí disponibles. Con 29916 registros es aún un número suficiente para realizar un análisis de regresión logística y de clasificación con XGBoost.

Variables a correlacionar con la percepción De las originales referidas al tiempo en que se produce el terremoto *Fecha* y *Hora* se trabajará solo con la generada a partir de la segunda *Hora_decimal* obviando la primera pues no es esperable que la percepción de un sismo por parte de la población. De las originales referidas a la magnitud y profundidad del sismo se trabajará con la generada a partir de la primera *proxy_amplitud* y la segunda sin modificación. Quedan así un total de tres variables a trabajar en el modelo de clasificación del estado de *Percibido*: *Hora_decimal*, *Profundidad* y *proxy_amplitud*. Con la función `cor` se verificó que la covarianza de *Percibido* con *Hora_decimal* es

casi nula, que cuanto más profundo es el terremoto es menos percibido y que `proxy_amplitud` tiene una correlación positiva, como era de esperarse, con la percepción de los sismos. Se resume esto en el cuadro 3.1.

Hora_decimal	Profundidad	proxy_amplitud	Percibido
-0.004116917	-0.142358740	0.165371903	1.000000000

Cuadro 3.1: La fila para la variable de percepción de la matriz de covarianza entre las variables de los sismos a analizarán muestra una débil correlación positiva con la variable que depende de la amplitud de las ondas sísmicas y una negativa con la profundidad del terremoto en concordancia con las expectativas lógicas que pueden tenerse sobre el fenómeno.

3.6. Preprocesamiento

Las evaluaciones sobre la calidad de los modelos de clasificación generados se realizarán sobre un subconjunto de ensayo (test) del 20 % de los datos de la provincia de San Juan, el resto se utilizará para el entrenamiento (train).

Escalamiento Previo a la partición (splitting) se realiza un escaleo uniforme sobre todo el conjunto de datos (scaling) de las tres características numéricas. El objeto de hacer esto previo a un ajuste lineal es que los coeficientes de la regresión sean comparables entre sí, con lo que la convergencia será más rápida y estable. Para esto se hace uso de la función ‘scale’ de la biblioteca ‘caret’ para generar `sismos_SJ_escalado`

Partición con estratificación Dado el fuerte desbalance de la clase Percibido comentado en la sección 3.3, ante una división de los datos en subconjuntos entrenamiento y prueba estocástica está el riesgo de que el subconjunto de prueba quede con muy pocos casos positivos y no sea representativo de la distribución de la clase en el conjunto de datos. Para evitar esto se realiza una división estratificada de los datos en subconjuntos de entrenamiento y prueba usando la función `CreateDataPartition` que indicó los índices para generar los conjuntos de datos `entrenamiento_SJ` y `ensayo_SJ` este último con un número aún adecuado para su función de 5982 registros.

Desbalance en entrenamiento Para contrarrestar el desbalance en la clase de clasificación se utilizará la técnica de sobremuestreo de la clase minoritaria que genera nuevos casos sintéticos de la clase minoritaria a partir de los existentes. Para esto se utiliza la función `ovun.sample` de la biblioteca `ROSE` que genera un conjunto de datos de entrenamiento con un número de casos de la clase minoritaria igual al de la clase mayoritaria. Se generó así un nuevo conjunto de datos de entrenamiento `entrenamiento_SJ_balanceado` con 23934 registros.

3.7. Descripción de las métricas de evaluación de los modelos (si corresponde)

3.8. Descripción de los métodos estadísticos utilizados (si corresponde)

3.8.1. Regresión logística para la predicción binaria

La siguiente es un resumen del material de la novena clase de la asignatura *Enfoque estadístico del aprendizaje* titulada *Regresión Logística* elaborado por Juan Barriola, Azul Villanueva y Franco Mastelli.

Un modelo de regresión lineal con coeficientes β_{α_i} para cada variable X_i que busque la probabilidad de una dependiente binaria $P(Y)$

$$P(Y) = \beta_0 + \sum_{j=1}^p \beta_j X_j,$$

no presentaría un punto de corte claro para clasificar los datos en dos categorías. Para sortear esta dificultad se toma la salida de esta regresión como la variable dependiente de una regresión logística,

$$P(Y|X) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}},$$

lo que asegura un valor entre 0 y 1. De esta expresión puede arribarse a

$$\log \frac{P(x)}{1 - P(x)} = \beta_0 + \sum_{j=1}^p \beta_j X_j,$$

cuyo lado izquierdo es el logaritmo de los *odds* y se llama *logit*.

La función `glm` de la biblioteca `glmnet` permite crear un modelo lineal generalizado (Generalized Linear Model). Al igual que la función de modelo lineal `lm` toma como argumentos una *formula* y los *datos* pero también se debe especificar el argumento *family*: indicamos la distribución del error y la función link que vamos a utilizar en el modelo según la distribución que corresponde a la variable a predecir

- Binomial: link=logit
- Poisson: link=log
- Gaussiana: link=identidad

Como estamos trabajando con un fenómeno que suponemos tiene una distribución binomial, así lo especificamos en el parámetro *family*.

3.8.2. XGBoost para una predicción binaria

Esto es un placeholder para una sección aún vacía.

Capítulo 4

Resultados y discusión

4.1. Presentación y análisis de resultados obtenidos

4.1.1. Regresión logística

Múltiple sin interacción entre variables

El primer modelo ensayado es uno múltiple de regresión logística con las variables `Hora_decimal`, `Profundidad` y `proxy_amplitud` como predictores sin interacción entre ellos,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Hora_decimal} + \beta_2 \text{Profundidad} + \beta_3 \text{proxy_amplitud}, \quad (4.1)$$

La función `summary` produce un resumen de los resultados arroja para los coeficientes:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.311098	0.014614	-21.287	<2e-16 ***
Hora_decimal	-0.015398	0.013362	-1.152	0.249
Profundidad	-0.555176	0.010705	-51.863	<2e-16 ***
proxy_amplitud	0.167536	0.006263	26.750	<2e-16 ***

Esto indica que excepto el coeficiente β_1 para la `Hora_decimal`, todos son significativos ya que sus probabilidades de que no tengan esos valores y se cumpla la hipótesis nula $Pr(> |z|)$ para el estadístico $z = \frac{\beta}{\sigma_\beta}$ se indican como muy inferiores a 0,05.

4.2. Discusión de los resultados y su relevancia

4.3. Limitaciones y posibles mejoras

Capítulo 5

Conclusión

- 5.1. Resumen de los hallazgos principales
- 5.2. Conclusiones generales y su relación con los objetivos del trabajo
- 5.3. Aplicaciones y relevancia de los resultados

Bibliografía

- [1] C. M. R. Fowler. *The Solid Earth: An Introduction to Global Geophysics*. first. Cambridge University Press, 29 de jun. de 1990. 490 págs. ISBN: 978-0-521-37025-7. DOI: 10.1017/CB09780511819643. URL: <https://archive.org/details/solidearthintrod0000fowl> (visitado 16-06-2024).
- [2] ¿Qué es un sismo? Sistema Nacional para la Gestión Integral del Riesgo. 12 de nov. de 2018. URL: <https://www.argentina.gob.ar/sinagir/riesgos-frecuentes/sismos> (visitado 16-06-2024).
- [3] Saunders, J. K., Minson, S. E., Cochran, E. S., Bunn, J., Baltay, A. S., Kilb, D. y O'Rourke, C. «A Twist of PLUM: Low-Magnitude Earthquakes and Ground-Motion-Based Early Warning». En: 2021 Southern California Earthquake Center Annual Meeting, SCEC Contribution #11360. URL: <https://www.scec.org/publication/11360> (visitado 17-06-2024).
- [4] Sandra Vaiciulyte, David A. Novelo-Casanova, Allen L. Husker y Ana B. Garduño-González. «Population response to earthquakes and earthquake early warnings in Mexico». En: *International Journal of Disaster Risk Reduction* 72 (1 de abr. de 2022), pág. 102854. ISSN: 2212-4209. DOI: 10.1016/j.ijdr.2022.102854. URL: <https://www.sciencedirect.com/science/article/pii/S2212420922000735> (visitado 17-06-2024).
- [5] *Instituto Nacional de Prevención Sísmica*. Argentina.gob.ar. 28 de oct. de 2022. URL: <https://www.argentina.gob.ar/inpres> (visitado 17-06-2024).
- [6] Yi Hu, Wentao Wang, Lei Li y Fangjun Wang. «Applying Machine Learning to Earthquake Engineering: A Scientometric Analysis of World Research». En: *Buildings* 14.5 (mayo de 2024). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, pág. 1393. ISSN: 2075-5309. DOI: 10.3390/buildings14051393. URL: <https://www.mdpi.com/2075-5309/14/5/1393> (visitado 17-06-2024).
- [7] Christoph Molnar. «9.6 SHAP (SHapley Additive exPlanations)». En: *Interpretable Machine Learning*. 26 de mayo de 2024. URL: <https://christophm.github.io/interpretable-ml-book/shap.html> (visitado 18-06-2024).
- [8] Heather Bedle, Diana Salazar-Florez y Christopher R. H. Garneau. «Recognizing societal influences in earthquake geohazard risk perception with explainable AI while mitigating risks through improved seismic interpretation». En: *The Leading Edge* 41.11 (nov. de 2022), págs. 756-767. ISSN: 1070-485X, 1938-3789. DOI: 10.1190/tle41110756.1. URL: <https://library.seg.org/doi/10.1190/tle41110756.1> (visitado 18-06-2024).

- [9] Itzhak Lior y Alon Ziv. «The Relation Between Ground Motion, Earthquake Source Parameters, and Attenuation: Implications for Source Parameter Inversion and Ground Motion Prediction Equations». En: *Journal of Geophysical Research: Solid Earth* 123.7 (2018), págs. 5886-5901. ISSN: 2169-9356. DOI: 10.1029/2018JB015504. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2018JB015504> (visitado 18-06-2024).
- [10] Avoy Datta, Daniel J. Wu, Weiqiang Zhu, Michael Cai y William L. Ellsworth. «DeepShake: Shaking Intensity Prediction Using Deep Spatiotemporal RNNs for Earthquake Early Warning». En: *Seismological Research Letters* 93.3 (16 de mar. de 2022), págs. 1636-1649. ISSN: 0895-0695. DOI: 10.1785/0220210141. URL: <https://doi.org/10.1785/0220210141> (visitado 18-06-2024).
- [11] J. L. Kirschvink. «Earthquake Prediction by Animals: Evolution and Sensory Perception». En: *Bulletin of the Seismological Society of America* 90.2 (1 de abr. de 2000), págs. 312-323. ISSN: 0037-1106. DOI: 10.1785/0119980114. URL: <https://pubs.geoscienceworld.org/bssa/article/90/2/312-323/120521> (visitado 18-06-2024).
- [12] Camilla Lingjærde, Martin Jullum y Nikolai Sellereite. *shapr: Explaining individual machine learning predictions with Shapley values*. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/shapr/vignettes/understanding_shapr.html (visitado 16-06-2024).
- [13] *IC-datasets-docencia*. URL: <https://daniellaparada.github.io/IC-datasets-docencia/> (visitado 12-06-2024).
- [14] Daniela Parada. *IC-datasets-docencia - 4 Visualización*. URL: https://daniellaparada.github.io/IC-datasets-docencia/04_visualizacion.html (visitado 12-06-2024).
- [15] *Buscador de sismos*. Instituto Nacional de Prevención Sísmica. URL: http://contenidos.inpres.gob.ar/buscar_sismo (visitado 15-06-2024).
- [16] *ISO 8601-1:2019(en), Date and time — Representations for information interchange — Part 1: Basic rules*. International Organization for Standardization. 2019. URL: <https://www.iso.org/obp/ui/en/#iso:std:iso:8601:-1:ed-1:v1:en> (visitado 16-06-2024).
- [17] *Introduction to data.table*. The Comprehensive R Archive Network. 27 de mar. de 2024. URL: <https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html#1.%20Basics> (visitado 18-06-2024).
- [18] William L. Ellsworth. «Earthquake Magnitude: THE RICHTER SCALE (ML)». En: *The San Andreas Fault System, California*. Ed. por Robert E. Wallace. Vol. Professional Paper 15151. P. United States Geological Survey (USGS), 1991, pág. 177. URL: https://web.archive.org/web/20160425121745/http://www.johnmartin.com/earthquakes/eqsafs/safs_693.htm (visitado 14-10-2008).
- [19] *Acerca de tu ubicación*. Instituto Nacional de Prevención Sísmica. URL: <https://www.inpres.gob.ar/desktop/conoce.html> (visitado 17-06-2024).

Anexos (opcionales)

- 5.4. Código fuente utilizado en el análisis (link a un repositorio)
- 5.5. Tablas y gráficos adicionales
- 5.6. Otros materiales relevantes